PITCH SYNCHRONOUS ANALYSIS/SYNTHESIS USING THE WRLS-VFF-VT ALGORITHM

BY

KYOSIK LEE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1992

In His Love:

to Soojeong, my wife, Sang-yub, my son, and Erica-Ensol, my daughter,

and my mother and parents-in-law

for their love, faith, patience and encouragement.

ACKNOWLEDGMENTS

I would like to thank my advisor and committee chairman, Dr. Donald G. Childers, for his invaluable guidance, assistance as well as financial support in the Mind-Machine Interaction Research Center (MMIRC) throughout this research. This tremendous help has been invaluable to me in many aspects of my scholarly as well as personal life.

I would also like to thank Dr. Leon W. Couch, II, Dr. Fred J. Taylor, Dr. Jose C. Principe, and Dr. Howard B. Rothman for their invaluable time and interest in serving on my supervisory committee. I wish to thank all my colleagues in the MMIRC for the stimulating discussions and help in many ways.

I wish to thank Soojeong, my wife, and Sang-yub, my son, and Erica-Ensol, my daughter, for their love, faith and patience throughout the years. I am also deeply indebted to my mother, sisters, brothers-in-law, parents-in-law, and uncle-in-law for their love and encouragement.

Finally I want to dedicate all my glory to my God, Jesus Christ.

TABLE OF CONTENTS

	page
ACKNOWLEDGMENTS	iii
ABSTRACT	vii
CHAPTERS	
1 INTRODUCTION	1
1.1 Parametric Linear Source-Filter Model	1
1.1.1 Source Excitation Factors	2
1.1.2 Filter Factors	5
1.2 Research Issues	6
1.3 Previous Research and Limitations	7
1.3.1 V/U/M/N/S Classification	7
1.3.1.1 V/U/M/S classifier	8
1.3.1.2 Nasal detection	10
1.3.2 Determination of Fundamental Frequency (F0)	12
1.3.3 Estimation of the Glottal Volume Velocity	14
1.3.4 Estimation of the Parameters	19
1.4 Research Objectives	22
1.5 Description of Chapters	23
2 RESEARCH DESIGN	25
2.1 Overview of Research	25
2.2 Experimental Data Base	28
2.2.1 Subjects and Tasks for Analysis	28
2.2.2 Data Base for the Application of Analysis/Synthesis	29
2.2.3 Data Collection	30
2.2.4 Microphone Characteristics	33
2.3 Preprocessing of Data	34
2.3.1 Demultiplexing and Trimming the Data	34
2.3.2 Synchronization of Data	34
2.4 Electroglottograph (EGG)	36

3 WRLS-VFF-VT ALGORITHM 37

3.1 Introduction 37
3.2 Algorithm Description 38
3.2.1 Background for the WRLS Algorithm 38
3.2.2 Adaptive WRLS-VFF-VT Algorithm 43
3.2.3 WRLS-VFF-VT Algorithm with Input Estimation 47
3.2.4 Complexity 53

4 ANALYSIS I : V/U/M/N/S CLASSIFICATION AND PITCH PERIOD DETECTION 54

4.1 Pitch Detection Algorithms 54
4.1.1 SIFT and Modified SIFT Algorithms for the Pitch Detection 57
4.1.2 EGG based Algorithm 60
4.1.3 LP error based Algorithm 64
4.1.4 VFF based Algorithm 66
4.1.5 Performance Evaluation 70
4.1.6 Summary 73
4.2 One-Channel Five-Way Classification 73
4.2.1 Introduction 74
4.2.2 Features 75
4.2.2.1 V/U/M/S decision 75
4.2.2.1.1 Methodology for spectral-domain features 78
4.2.2.1.2 Statistical properties of features 81
4.2.2.2 Nasal/nonnasal decision 81
4.2.2.2.1 Statistical properties of features 100
4.2.3 Pattern Classification 101
4.2.3.1 Vector quantization 103
4.2.3.1.1 Introduction 103
4.2.3.1.2 Distortion measures 106
4.2.3.1.3 Codebook design 109
4.2.3.1.4 Selection of codebook size 111
4.2.3.2 Neural network 111
4.2.3.2.1 Multi-layer perceptron 112
4.2.3.2.2 Network structure 114
4.2.3.2.3 Data normalization 115
4.2.3.2.4 Complexity requirements 117
4.2.3.2.5 Training and testing 119
4.2.3.2.6 Criteria to stop the training 122
4.2.3.3 Decision tree method 125
4.2.3.3.1 V/U/M/S classification 125
4.2.3.3.2 Nasal/nonnasal classification 136
4.2.4 Result 137

4.3 Pitch Synchronous V/U/M/N/S Classification 142
4.4 Summary 150

5 ANALYSIS II : ADAPTIVE FORMANT TRACKING AND GLOTTAL INVERSE FILTERING (GIF) USING CLOSED PHASE WRLS-VFF-VT 152

5.1 Introduction 152
5.2 Estimation the Speech Parameters Based on the WRLS-VFF-VT 154
5.2.1 Adaptive Formant Tracking using Closed Phase WRLS-VFF-VT 155
5.2.2 Closed phase WRLS-VFF-VT algorithm 155
5.2.3 Performance Evaluation of WRLS-VFF-VT Algorithm 156
5.2.3.1 The generation of synthetic speech signal 157
5.2.3.2 Experimental results 158
5.2.4 Summary 169
5.3 Glottal Inverse Filtering 170
5.3.1 Background for the Glottal Inverse Filtering 171
5.3.2 Glottal Wave Estimation 171
5.3.3 Glottal Inverse Filtering 172
5.3.3.1 GIF using WRLS-VFF-VT 176
5.3.4 Voice Source Models 178
5.3.4.1 Fant's model 180
5.3.4.2 LF-model 183
5.3.5 Modeling of the Glottal Flow Waveform 187
5.3.5.1 Measurement of model parameters 187
5.3.6 Experimental Results of GIF and the LF model matching .. 190
NO TAG Comparison of Different GIF Methods 193
NO TAG Summary 199

6 SYNTHESIS AND APPLICATION 203

6.1 Synthesis Strategies and Experiment results 203
6.1.1 Voiced Sounds 206
6.1.2 Unvoiced Sounds 210
6.1.3 Mixed Excitation Sounds 219
6.1.4 Summary 222
6.2 Automatic Classification of Different Voice Types 226
6.2.1 Glottal Waveform Characteristics 228
6.2.2 Results of Simple Statistics 231
6.2.3 Automatic Classification results 232

7 CONCLUSIONS AND DISCUSSION 236

7.1 Summary 236

7.2 Application 238
7.3 Future Work 239

APPENDIX DIFFERENT ADAPTIVE FILTERING METHODS FOR SPEECH ANALYSIS 241
A1 Summary 241
A2 Introduction 241
A3 System Configuration 243
A4 Analysis of Experiment Results 245
A4.1 Comparison among DEGG, VFF and Differentiated Adaptive - μ 245
A4.2 Comparison of the Estimation Ability among WRLS-VFF with AR Model, ALMS with Gamma Model and NLMS with AR Model 245
A4.3 Comparison of the Computational Complexity among WRLS-VFF with AR Model, ALMS with Gamma Model and NLMS with AR Mode 254
A5 Conclusion 255

REFERENCES 257

BIOGRAPHICAL SKETCH 276

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

PITCH SYNCHRONOUS ANALYSIS/SYNTHESIS SYSTEM USING WRLS-VFF-VT ALGORITHM

By

Kyosik Lee

December 1992

Chairman: Dr. D.G. Childers
Major Department: Electrical Engineering

In this study, the pitch synchronous and asynchronous analysis algorithms were developed for the formant synthesizer.

New pitch detection algorithms, which are Variable Forgetting Factor (VFF) based, EGG based, and LP error based methods, are introduced. Our pitch detectors are very reliable in quasi-periodic as well as in aperiodic speech signals. The "pitch smearing" effect inherent in speech signal based methods is avoided, and pitch values are available on a period by period basis. Furthermore, the locations of the glottal closing instants allow the isolation of individual periods of the speech waveform from closure to closure.

A fairly general framework based on a pattern recognition approach to Voiced/Unvoiced/Mixed/Nasal/Silent (V/U/M/N/S) classification has been described in which a set of measurements is made on the interval being classified, and Vector Quantization (VQ), Neural Network (NN), and decision tree classifiers are used to select the appropriate class. The work constitutes a demonstration that the V/U/M/N/S classification can be made with reasonable accuracy. We have also developed a

pitch-synchronous V/U/M/N/S classification method. The method provides the information useful for fine segmentation of the speech signal and also provides a classification rate as good as the pitch-asynchronous algorithm discussed in this study. The pitch synchronous analysis method will be used in a complete analysis/synthesis system.

A new weighted recursive least squares algorithm with a variable forgetting factor and with a variable threshold (WRLS-VFF-VT), which is for estimating ARMA parameters, input pulse train, and input white noise at the same time so that formants and antiformants of speech can be correctly estimated, was developed.

A new glottal inverse filtering technique using the WRLS-VFF-VT method was proposed. This method uses the three algorithms for the detection of the glottal closing instants using the VFF, LF error, or EGG signals depending on the characteristics of the input signal. It can provide reliable glottal v-v waveform estimates automatically for the normal speech, synthetic speech, and the pathological speech signals. The proposed algorithm was compared with the two-channel CPC method for analyzing different types of speech signals.

In summary, the major contributions of this dissertation were the development of new speech analysis techniques for designing new systems for producing natural sounding synthetic speech with a desired voice quality. Our complete analysis/synthesis system using our method is being used for voice conversion, synthesis of high-quality speech, and synthesis of various voice types.

CHAPTER 1
INTRODUCTION

In the production of speech, sound is generated in the vocal system either by the vibration of the vocal cords or by the creation of turbulent air flow at a constriction. The sound produced is spectrally shaped by the transmission characteristics of the vocal tract which consists of the pharynx, the nasal tract and oral cavity. Speech production can be modeled as a slowly time-varying linear system that is excited by a source, which is a quasi-periodic pulse signal for voiced speech or as a flat spectrum random signal for unvoiced speech.

Speech synthesis is the process of producing an acoustic signal by controlling and updating the speech production model with an appropriate set of parameters. The model parameters can be obtained either by the analysis of real speech signals or by the analysis-by-synthesis procedure. If the model is sufficiently accurate and its parameters are accurately estimated, the resulting output of the model is comparable to natural speech.

1.1 Parametric Linear Source-Filter Model

A primary goal of speech analysis-synthesis is to gain an understanding of human speech production. The basic tool employed in this scientific exploration is a conceptual model of the vocal source and tract, for which various forms exist. Perhaps the common goal for the various models is the production or reproduction of

high-quality speech. Presumably, the better our analysis-synthesis models, the better the quality of the synthetic speech.

Acoustic and source-filter modeling have been successful, in that formant and LPC synthesizers can produce highly intelligible speech that sounds similar to that of a human speaker (Holmes, 1973; Klatt, 1987). However, high quality synthesized speech is analysis-dependent (Childers and Wu, 1990).

Several factors that appear to affect the quality of synthetic speech produced by analysis-synthesis are: (1) the filter factors, which include formant locations and bandwidths (or number of poles and their positions) and (2) the excitation wave shape, which may include source-tract interaction (Childers and Wu, 1990).

1.1.1 Source Excitation Factors

From a review of the literature three excitation factors appear to be important for synthesizing speech. We classify these factors as excitation timing, excitation waveshape, and source-tract interaction (Childers and Wu, 1990).

1.1.1.1 Excitation timing factors

Under this category we include the estimation of speech segments produced by either voiced (V), unvoiced (U), or mixed (M) excitation. (Experiments show that high quality synthesis requires mixed excitation for synthesis of the voiced fricatives (v(vote), th(then), z(zoo), z(azure)), the human production of which involves the vibration of the vocal cords in conjunction with a turbulent air flow at some point of constriction.) Silent (S) intervals must also be estimated by the analysis procedure. In many synthesis schemes, an accurate decision regarding V/U/M/S intervals is the most demanding requirement because even a single misclassified frame can lead to perceptible distortion in the synthesized speech, especially if a voiced segment is misclassified as U/S or vice-versa. This is due to the improper selection of the excitation

for LPC or using the wrong branch of the formant synthesizer, e.g., cascade instead of parallel. In order to allow a mixed source in an analysis-synthesis system, the excitation for a segment of speech must be identified as voiced, unvoiced, or a combination of voiced and unvoiced.

The measurement of the fundamental frequency of voicing (F_0) is another factor we include under excitation timing factors. Accurate pitch estimation is essential since errors in this measurement can have a significant effect on speech analysis at later stages. For example, the quality of synthesized speech is very highly dependent on accurate pitch estimates (Childers and Wu, 1991). Deviations of F_0 greater than 1% can be perceived by experienced listeners (Laver, 1980).

Natural voiced speech has period-by-period interval perturbations (jitter) and amplitude variations (shimmer). These factors, especially jitter, must be measured and used in speech reproduction if a high-quality voice replication is to be achieved. Pitch perturbations have been found useful for characterizing pathological and normal voices (Lieberman, 1961, 1963). There may be considerable variations over several pitch periods. Therefore, if possible, pitch synchronous analysis should always be used (Childers and Wu, 1991; Krishnamurthy and Childers, 1986).

1.1.1.2 Excitation waveform factors

The shape of the glottal pulse varies greatly from speaker-to-speaker for different speaking tasks (Monsen and Engebretson, 1977) and affects the quality and naturalness of synthetic speech (Rosenberg, 1971; Holmes, 1973, 1983; Klatt, 1987). Over the years many models have been proposed to generate an excitation signal that resembles the glottal volume-velocity (v-v) waveform or the prediction analysis residue signal (Rosenberg, 1971; Holmes, 1973, 1983; Klatt, 1987; Lee and Childers, 1989; Pinto et al., 1989).

Various types of glottal v-v waveforms have been used for exciting a synthesizer. Single-pulse excitation has been thought for some time to introduce "buzziness", making the synthetic speech sound unnatural (Sambur et al., 1978; Naik, 1983). One approach to solve this buzziness problem is to use multi-pulse excitation (Atal and Remde, 1982) and stochastic coding (Schroeder and Atal, 1985). A simpler approach is to use three impulses (positive, negative, positive). The spacing between the first impulse and the second is related to the glottal opening interval and the spacing between the second impulse and the third represents the glottal closing interval. The rationale for this triple impulse pulse excitation is as follows. Two poles of the LPC model can be used to represent the source. These poles act as integrators in the time-domain, which when applied to the three-impulse excitation yields a triangular-shaped waveform, similar in shape to the volume-velocity waveform. This excitation waveform is preferred by listeners of synthetic speech over single-pulse excitation, yielding a "crisper" quality with less buzziness (Childers et al., 1989). Other excitation waveforms have been suggested by Fant (1979), Ananthapadmanabha (1984), and Fant et al. (1985).

1.1.1.3 Source-tract interaction

We know from the experiments of Rothenberg (1981), Holmes (1973) and Sambur et al. (1976) that the shape of the excitation waveform affects the perceived quality of synthesized speech. The shape of the glottal volume-velocity waveform is affected by both the source and the vocal tract. One may often observe a "hump" in the rising portion of the glottal volume-velocity waveform of male speech obtained by using the glottal inverse filtering. This hump or ripple is presently thought to be due to source-tract interaction, i.e., the interaction of the source with the first formant (Fant, 1979).

Source tract interaction has been conjectured to be important for synthesizing high-quality, natural-sounding speech (Rothenberg, 1981; Childers et al., 1983). Source-tract interaction is the result of intra-pitch period phenomena that occur in speech production. This effect can be inserted into synthesis models (Childers et al., 1983; Yea and Childers, 1983; Wong, 1991) by using a "glottal area" excitation function to control the time-varying glottal impedance in an equivalent circuit of the vocal system. The output of the circuit is a time-varying "glottal volume-velocity" function that includes the effect of source-tract interaction. This glottal volume-velocity function is then used as the glottal excitation for the formant synthesizer.

1.1.2 Filter Factors

Accurate determination of the resonant frequencies of the vocal tract (i.e., formant frequencies) in various articulatory configurations is of interest in the analysis-synthesis of speech. For vowel sounds, in particular, the formant frequencies (especially the first and second formants) play a dominant role in determining which vowel is produced by a speaker and which vowel is perceived by a listener. Formant bandwidths, on the other hand, affect speech quality and, in any synthesis procedure, must be selected properly to achieve natural-sounding speech. In speech analysis, the formant bandwidths may affect the accuracy attainable by formant-frequency locating procedure.

The acoustic structure of nasal consonants has long been predicted by the acoustic theory of speech production (Fant, 1960; Fujimura, 1962; Flanagan, 1972). The presence of a side-branching resonator (the blocked oral cavity) will introduce an antiresonance (zero) in the spectrum of nasal consonants. Theoretically, the antiresonance can be used to identify the place of articulation of nasal consonants because the frequency of the antiresonance is determined by the dimension of the

side-branching resonator. Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator and anti-resonator into the cascade vocal tract model in a formant synthesizer. The Klatt synthesizer includes an additional resonance-antiresonance pair for synthesizing nasals. It is necessary to identify nasalized segments in the recorded speech in order to decide when this branch should be activated. Another purpose of nasality detection is the correction of the all-pole estimates of the formant frequencies and bandwidths; it is known that the presence of zeros in the spectra of nasal speech tends to move the formants upward in frequency and to increase their bandwidths. Therefore, to synthesize a high quality speech from analysis parameters for the nasal consonants, extending the V/U/S or the V/U/M/S decision to V/U/M/N/S decision is needed.

Recently, it has been shown that representation of the process of speech production by an autoregressive moving-average (ARMA) process is superior to that based on an AR process in the spectral analysis of most of the consonants, especially the nasal and fricative consonants (Steiglitz, 1977; Morikawa and Fujisaki, 1982; Atal and Schroeder, 1978; Miyanage et al., 1982, 1986).

1.2 Research Issues

The quality of synthesized speech is affected by the excitation and the filter used for synthesis. For the excitation, one factor that needs to be considered is excitation timing characteristics, e.g., whether or not the excitation is voiced, unvoiced or mixed. In addition, pitch detection and jitter measurements are important. If possible pitch synchronous analysis should always be used. Another factor is the shape of the excitation waveform and its spectrum. A third factor is source-tract interaction, which remains difficult to measure but can be modeled as part of the excitation waveform. The measurement of each of these factors is dependent on several analysis

techniques, e.g., V/U/M/S measurements, pitch detection, jitter measures, and glottal inverse filtering for volume-velocity estimation.

For the filter factors, accurately tracking the parameters such as vocal tract resonances/antiresonances (formants/antiformants) and their bandwidths, which may change rapidly for different sounds or the transition from one sound to another, is necessary for the quality of synthesized speech.

Therefore, in the context of the linear source-filter model based on the parametric representation, basic analysis procedures for producing excellent quality synthesized speech are as follows:

for the source,

1) classification of the speech signal into voiced, unvoiced, mixed, nasal, and silent (V/U/M/N/S) segments.

2) determination of the fundamental frequency (F_0) and the glottal closed instants of vocal fold vibration in the voiced segments.

3) estimation of the glottal volume velocity (v-v) for the voiced segments.

and for the filter,

4) estimation of the time-varying ARMA parameters.

1.3 Previous Research and Limitations

1.3.1 V/U/M/N/S Classification

Segmentation of speech is required in many areas of speech processing and coding such as speech interpolation, vocoding, and speech recognition. The accuracy of segmentation is one of the important factors that affects the overall system performance directly.

There have been many studies on acoustic segmentation of speech with results ranging from simple speech detection algorithms to a V/U/M/S classification algorithm.

1.3.1.1 V/U/M/S classifier

Atal and Rabiner (1976) suggested a pattern recognition approach to V/U/S classification based on five measurements or features: 1) the zero-crossing rate, 2) the speech energy, 3) the correlation between adjacent speech samples, 4) the first predictor coefficient formed by a 12-pole linear predictive coding (LPC) analysis, and 5) the energy in the prediction error. These five measurements were combined using a non-Euclidian distance metric to have a reliable decision. A final correct classification rate of 96.6% was reported. Unfortunately, there was no comment on how the voiced fricatives in the input sentences were handled and the data set seemed to be too small to produce a generalizable result.

Siegel and Bessey's V/U/M classifier (Siegel and Bessey, 1982) used the following features: 1) speech energy, 2) normalized autocorrelation coefficient at unit sample delay, 3) linear prediction error, 4) the first linear predictor coefficient, 5) zero crossing rate, 6) the ratio of energy in the signal above r(H) Hz to that below r(L) Hz (in fact, three ratios were used). The final recognition rate of 94.0% was claimed. There are four major disadvantages to this classifier. 1) It accepts only the speech part of input sequences, so an operator was needed to manually eliminate silent intervals from the input sentences before they were processed. As a result, the performance of the system was greatly improved, but it made the system less attractive due to the difficulty of providing endpoint information essential to isolated word recognition (IWR) systems. 2) The authors failed to describe clearly how mixed excitation intervals were identified manually. 3) The absence of silent interval detection capability can be critical in some applications, such as automatic control of the excitation mode in speech

synthesis and codeword generation in IWR systems. 4) The final overall error rate was not given.

Larar's V/U/M/S classifier (Larar, 1985) was a two-channel method, using both speech and EGG (electroglottography) as its input signals. The features used were 1) the zero crossing rate of the speech signal, 2) the energy of the EGG signal, and 3) the energy of the speech signal. The V/M/U/S classification was heavily dependent on the presence of the relatively high EGG energy in the frame. However, the use of the energy of the EGG signal in order to detect vocal fold vibration can deteriorate the system significantly when there exists a relatively large low frequency fluctuation in the signal. The data set can be seen to have two weak points: 1) the number of silent frames is too large, 69.8% of total frames, to declare that the 95.45% correct rate is a generalizable one, and 2) the number of mixed frames is too small to assert the final mixed frame identification rate, 87.5%, as a reliable one.

Hahn (1989) presented two different algorithms for automatically classifying speech into four categories. The algorithms employed information from either two-channels (speech and EGG) or one-channel (speech only). An overall correct classification rate of 98.7% was achieved for the two-channel algorithm, when judged against skilled manual classification. For the one-channel algorithm, the overall correct rate was slightly less, at 96.9%. The features that were selected for one-channel four-way classification algorithm were 1) speech energy, 2) zero crossing, 3) level crossing rate, 4) zero crossing rate of the differentiated speech signal, and 5) spectral distribution. This algorithm has one serious and several minor drawbacks. The serious drawback centers on the requirement that the clear-cut voiced-unvoiced segments need to be classified initially to determine the average and standard deviation before the criterion for a mixed excitation can be applied. A mixed excitation segment in which the unvoiced component was stronger than the voiced component would never be tested for a mixed excitation. There are three minor drawbacks. First, the single

feature considered for use in making the mixed classification was the spectral energy distribution, which was not sufficient to discriminate between mixed and other excitation segments. This was shown by only a 69.9% correct identification of mixed excitation frames. Second, several threshold values calculated in a fixed frame make it difficult to apply this algorithm to the variable frame-size analysis for a pitch-synchronous synthesis/analysis of speech. Third, the complicated decision-tree structure that was used makes it difficult to modify and update the algorithm for new data.

1.3.1.2 Nasal detection

It is widely accepted that certain spectral characteristics of nasal murmurs act as acoustic cues. Phoneticians suggest the relevance of the following distinctive spectral traits of nasal murmurs (Fujimura, 1962):

- A low first nasal formant (N1) at 250-300 Hz, with higher intensity than the upper formants

- A nasal antiformant (NZ), varying in frequency with place of articulation;

- A set of weak formants (N2, N3, N4,...) at 300-4000 Hz, with large bandwidth(bw) values;

- An overall lower intensity level than vowels.

Other less important cues are

- Nasalization effects upon the neighboring vowels;

- Nasalized releases;

- F1 transitions which are less pronounced than for nonnasal stops.

Mori et al. (Mori, R. D., Gubrynowicz and Laface, 1979) proposed an algorithm for the recognition of two classes of nasals, the bilabial /m/ and alveolar /n/. The algorithm is based on a syntactic pattern recognition approach. A set of rules in the

form of fuzzy relations for generating hypotheses about intervocalic nasal consonants was inferred from the experiment. Coarticulation effects were accounted for in generating the recognition rules to improve the classification of nasal sounds and substantially better results were obtained than in the previous approaches.

Demichelis (1982) proposed an algorithm for nasal recognition which subsequently gave better results than any other reported methods. In this system, relationships between acoustic cues and phonetic and phonemic interpretation were established by a decision method based on a possibility distribution. The algorithm consisted of two subsystems; the high level subsystem classified the nasal sounds as one category, and the low level subsystem then performed the identification of individual nasal phonemes.

A high order linear prediction (LP) analysis was used to estimate the spectrum of speech by Yea and Childers (1983). The use of a high order LP was motivated by the fact that extra pole-zero pairs exist in the vocal tract transfer function for nasalized sounds. Also, the formants were more closely spaced for nasal consonants since the nasal tract is longer than the oral tract. Yea's algorithm accurately identified the nasal consonants but did not perform as well in identifying nasalized vowels. Errors tended to occur at the end of voicing. No overall statistical error rate was calculated since the test data set was not extensive .

Kurowski and Blumstein (1987) used a simple method for representing spectral change at the nasal - vowel boundary, in which they examined the proportion of change in energy in the region of Bark 5-7 to the change in the region of Bark 11-14 across two spectra. This method , which did not use predetermined frequency bands for computing proportions of spectral change, involved simply subtracting the vowel spectrum from the murmur spectrum for each spectrum pair, and calculating four parameters: (1) the frequency at which the maximum of the difference spectrum occurs, (2) the value (in dB or phons) of the maximum of the difference spectrum, (3) the

frequency at which the minimum of the difference spectrum occurs, and (4) the value (in dB or phons) of the minimum of the difference spectrum.

1.3.2 Determination of Fundamental Frequency (F_0)

Accurate and reliable measurement of the pitch period of a speech signal from its acoustic pressure waveform alone is often exceedingly difficult for four reasons. The first reason is that the glottal excitation waveform is not a perfect train of periodic pulses. The second is the interaction between the vocal tract and the glottal excitation. The third is the inherent difficulty in defining the exact beginning and end of each pitch period during voiced speech segments. Finally, the fourth reason is distinguishing between unvoiced speech and low-level voiced speech. When transmitting speech through a telephone system, additional complications occur when one is faced with the problem of pitch extraction of speech.

Basically, a pitch detector is a device which makes a voiced-unvoiced decision, and provides a measurement of the pitch period during periods of voiced speech. However, some pitch detection algorithms determine only the pitch period during voiced segments of speech and rely on some other technique for the voiced-unvoiced decisions.

Pitch detection algorithms can roughly be divided into the following three broad categories.

1) A category which utilizes principally the time-domain properties of speech signals.
2) A category which utilizes principally the frequency-domain properties of speech signals.
3) A category which utilizes both the time- and frequency-domain properties of speech signals.

Time-domain pitch detectors operate directly on the speech waveform to estimate the pitch period. For these pitch detectors the measurements most often

made are peak and valley measurements, zero-crossing measurements, and autocorrelation measurements. The basic assumption that is made in all these cases is that if a quasi-periodic signal has been suitably processed to minimize the effects of the formant structure, then simple time-domain measurements will provide good estimates of the period.

The class of frequency-domain pitch detectors use the property that if the signal is periodic in the time domain, then the frequency spectrum of the signal will consist of a series of impulses at the fundamental frequency and its harmonics. Thus simple measurements can be made on the frequency spectrum of the signal (or a nonlinearly transformed version of it as in the cepstral pitch detector) to estimate the period of the signal.

The class of hybrid pitch detectors incorporates features of both the time-domain and the frequency-domain approaches to pitch detection. For example, a hybrid pitch detector might use frequency-domain techniques to provide a spectrally flattened time waveform, and then use autocorrelation measurements to estimate the pitch period.

As a result of the numerous difficulties in pitch measurement, a wide variety of sophisticated pitch detection methods have been developed. Most of these studies have focused on pitch determination in stationary, quasi-periodic speech signals. Nonstationary, aperiodic signals, i.e. exhibiting occasional time and/or amplitude irregularities between successive periods of glottal excitation, have traditionally been disregarded and considered as reflecting pathological voice phenomena. These are of no immediate concern to normal speech processing applications. Standard pitch detection and pitch estimation algorithms usually fail to correctly identify patterns of aperiodic voice excitation. Most of the standard pitch detection algorithms provide only an average pitch estimate over a number of periods, because they rely on the similarity of the speech signal over adjacent pitch periods. Thus, if a large number of

pitch periods are contained in an analysis segment the estimated pitch value is a "smeared" or average value for the segment.

1.3.3 Estimation of the Glottal Volume Velocity

The determination of the glottal wave has been a target of research during the past several decades. To compute the exact waveform of the excitation of the vocal tract is not easy. The reason for this is, first, that the monitoring and the measuring of the characteristics of the vibrating vocal cords are difficult. Second, if the analysis of the glottal flow is based on the acoustic signal, the characteristics of the filtering part, that is the vocal tract, are complicated. As a consequence of the complexity of the analysis, different kinds of methods have been developed for determination of the true glottal wave.

Miller (1959) developed a glottal wave analysis method that is based on analog circuits. By performing a careful spectrographic analysis the frequency position of the first formant was initially determined. The inverse filter circuit was further tuned so that the glottal pulse estimate would disappear during the closed period of the vocal cord wave. Results of this study show that for male voices the glottal wave is characterized by a clear closure period, a gradual rising and more abrupt closure. The glottal wave became a little more triangular at higher pitches and by increasing the stress the pulse form changes from almost sinusoidal to the form having a clear closure. For female voices the opening and closing of the glottis were reported to be of about equal steepness. In the paper by Mathews et al. (1961) a pitch synchronous analysis method was presented. By computing the Fourier analysis for one period of a voiced sound and by approximating the result of the transformation by a pattern of zeros and poles, the source and the vocal tract were separated. The zeros of the spectrum were used to estimate the open-to-closed time for the glottis.

Rosenberg (1971) developed a pitch synchronous method where the formants were modeled with a "trial spectrum" that was computed from the Fourier transform of one pitch period. A residual signal was obtained by eliminating the effect of the formants on the speech signal with inverse filtering. Finally, the excitation waveform was calculated by the inverse Fourier transform of the residual. The method was used to study the effect of the detailed structure and timing parameters of the excitation in the synthesis of speech.

A different method to compute the glottal volume velocity waveform was presented by Rothenberg (1973), where the use of a pneumotachograph mask was introduced. Many advantages of this technique over the previously developed methods were reported. The effect of fundamental frequency, subglottal pressure and breathy/unbreathy voicing on the shape of the glottal waveform was discussed.

A new, alternative method was presented by Sondhi (1975): the excitation of the vocal tract was obtained by eliminating the effect of the resonances by speaking into a reflectionless uniform tube. This real time method is said to be inexpensive and robust against noise.

Holmes (1976) used partial inverse filtering to produce waveforms representing single formants of voiced speech. His results confirmed the well known fact that the main formant excitation normally occurs at glottal closure. However, there is, frequently, evidence of additional excitation, not only at glottal opening and during the open phase, but also after closure.

Hunt et al. (1978) described an interactive digital inverse filtering system in which the advantages of analog and digital methods are combined to provide a facility with much greater convenience and power than either.

Wong et al. (1979) suggested a straight forward (but computationally expensive) approach for performing glottal inverse filtering from the acoustic speech waveform by analyzing the normalized linear prediction error sequence obtained by

calculation of the p-pole total linear predictive error on an M-point window of the speech waveform. Both the moment of glottal closure and opening can be determined from the normalized total squared error with proper choices of analysis window length and filter order. The window is moved through the speech waveform one point at a time and after energy normalization the total-error sequence represents a measure of the fit of a p-pole model to segments of the waveform. The total error is at a minimum (ideally, zero) during analysis of a completely closed phase segment. However, in cases of high frequency or breathy speech where the closed phase is of shorter duration, this method may not provide unambiguous local minimal ranges of the normalized error sequence that are necessary to indicate closed phase. This problem occurs when the duration of the closed phase is less than the length of the analysis window. To estimate the actual volume velocity waveform, the filter for a single period was chosen as that corresponding to the minimum normalized error.

A difficult problem in the closed phase covariance method is the determination of the opening and closing of the glottis. This problem has been studied in Ananthapadmanabha and Yegnanarayana (1979) using the so called Hilbert envelope of the windowed linear prediction residual and in Strube (1974) with the help of the determinant of the autocovariance matrix of speech.

A pitch asynchronous, two-step glottal wave analysis method was proposed in Matausek and Batalov (1980). In the first step the covariance analysis is applied to the preemphasized speech signal and the model for the vocal tract is formed. By deemphasizing and integrating the residual, the signal u(k) is obtained. In the second step a glottal model Hg(z) is identified by applying an iterative inverse filtering method to the signal u(k). Finally, the glottal wave estimate is obtained by time reversing the response of Hg(z) to the impulse train.

Fant (1982) has studied the covariation of flow parameters with voice intensity and pitch using analog inverse filtering. There appear to be two modes available to

produce an intensity increase. One is a rise in the overall scale factor of glottal flow pulses which is a main consequence of increased subglottal pressure. The other is an adduction of the vocal folds physiologically induced by a medial compression, which may increase the steepness of the closing branch of glottal pulses while maintaining or even reducing the amount of air contained in a single pulse. For F0-variations, the flow amplitude shows a maximum around 115 Hz and then decreases in inverse proportion to F_0. The flow derivative, indicative of formant amplitudes, shows a maximum at F_0=118 Hz and then exhibits a fall rise contour indicating increased efficiency at higher pitch.

Using electroglottographic (EGG) signals together with the closed-phase inverse filtering technique is discussed in Veeneman and BeMent (1985). Reliable glottal wave results were obtained when the speech of 30 adults including both normal and pathological female and male subjects were tested. In Krishnamurthy and Childers (1986) a two channel technique using both EGG and speech was introduced. The method can be used not only in glottal wave analysis but also in classifying speech segments and in F0 estimation.

A number of limitations can be raised for closed glottal interval LPC analysis. The first is that the closed glottal interval is difficult to locate. For a real-time voice source analysis in a vocoder, this is a valid limitation, but in clinical speech situations, human intervention to identify closed glottal intervals with the aid of an electroglottograph (EGG) is not an unrealistic requirement. However, for the two-channel method, a second additional channel of data needs to be collected and processed. The two-channel method often fails when there is incomplete or no glottal closure. A more serious limitation is that with high pitch voices, the closed glottal interval contains too few samples to permit an effective least-squares determination of the AR coefficients. With some voice types, the vocal folds may not fully close, and these voice types are of clinical interest.

Lee and Childers (1988) used a two-pass method to perform the inverse filtering on the speech signal. In the first pass, the locations of the main pulses of the LP error signal were identified. Then, using these main pulses as indicators of glottal closure, a "psuedo closed phase" was selected as the analysis interval for a pitch-synchronous covariance LP analysis to estimate the vocal tract filter, which in turn was used to obtain the desired glottal volume-velocity waveform.

Ting (Ting, 1989; Ting and Childers, 1990) used a WRLS-VFF algorithm to get a continuous glottal volume-velocity waveform. The performance of three GIF methods(two-pass, two-channel and WRLS-VFF) were evaluated by analyzing different types of speech signals such as modal voice (medium pitch) and falsetto voice(high pitch). For the modal voice a closed phase interval was present and all three methods were able to extract the glottal volume-velocity waveform from the speech. For the falsetto voice analysis, due to the high pitch and incomplete glottal closure, the two-channel method failed to analyze the signal. The two-pass method could estimate the glottal volume velocity waveform for a only few intervals. In contrast, the WRLS-VFF method, as long as the pitch period is longer than twice the filter order, gave a good estimate of the glottal volume-velocity waveform.

A WRLS-VFF method (Ting and Childers, 1990) has several limitations. The first is that this method uses an off-line threshold value with the LP residual signal or the EGG signal for the detection of the pitch period and of the input excitation starting point. Due to the high variation of pitch and glottal closure instant for a sentence, an automatic GIF can not be performed. The second limitation is that the fixed threshold value for classifying the input source estimation often contributes to errors in the next frame. The final objection is that there is no preprocessing techniques such as V/U/M/N/S classification to be used to estimate the model type and model order before executing the WRLS-VFF algorithm.

Generally, glottal inverse filtering methods do not consider nasal consonants, nasalized vowels, and mixed excitation sound. In the case of nonnasalized vowels produced by normal male subjects, closed glottis LPC inverse filtering can be quite successful in terms of providing estimates of the vocal tract formants and the glottal volume velocity waveform that are consistent with acoustic theory. This is to be expected because the underlying assumptions of the closed glottis LPC technique are consistent with the acoustic model of these vowels. On the other hand, nasals, nasalized vowels, mixed excitation sound or speech corrupted by noise, have pole-zero spectra that do not fit the all-pole spectrum model of LPC analysis. As a consequence, there are inherent modeling errors in inverse filtering of the nasalized vowels using any of the LPC techniques.

1.3.4 Estimation of the Parameters of the Time-Varying Vocal Tract Filter

The estimation and tracking of formants and their bandwidths have long been recognized as important adjuncts to speech and speaker recognition and speech synthesis. Despite the popularity of the McCandless (1974) formant tracker, formant tracking may still be greatly improved. Several factors to consider when extracting the formants from speech signals are: 1) the periodicity of the excitation, 2) the duration and placement of the speech analysis frame, which affects the estimation of the formants and the influence of source-tract interaction, 3) the influence of the fundamental frequency of voicing (F_0) when it is near the first formant, 4) the effect of spectral notches caused by nasals or other consonant sounds, and 5) rapid variations of the formants that may occur in consonant-vowel transitions or diphthongs.

LPC techniques attempt to model the vocal tract and can provide an estimate of the envelope of the speech spectrum (Makhoul, 1976; Markel et al., 1976; Rabiner et al., 1978). However, frame-based LPC analysis methods cannot reduce the effect of source-tract interaction (Childers et al., 1984) nor can they track rapid formant

changes. The deficiencies of these techniques are caused by windows that average the data over several excitation epochs. Pitch synchronized LPC analysis, such as the closed phase covariance (CPC) method can reduce the effect of source-tract interaction (Krishanamurthy and Childers, 1986). However, the CPC vocal tract filter may, on occasion, be unstable yielding formant variations that cannot be tracked accurately (Ting and Childers, 1988; Ting, 1989). Cases where this may occur are 1) the fast transitions between vowels and consonants and 2) speech with short closed (or no closed) glottal intervals as may occur with female or child speech. Furthermore, another problem exists in that the all-pole model using LPC analysis does not estimate the antiformants present in nasals, nasalized vowels, or speech corrupted by noise (Pagano, 1974; Tierney, 1980).

Sequential adaptive methods with an autoregressive and moving average (ARMA) model offer an attractive alternate processing strategy for speech since they overcome some drawbacks of frame-based analysis methods (Friedlander, 1982; Morikawa et al., 1982). The ARMA methods provide accurate parameter estimates for a stationary process and also slowly track the model parameters. However, they lack the ability to reduce the influence of periodic pulse trains and source-tract interaction on parameter estimation. This is a problem because these methods estimate the excitation using the residual error, which is assumed to be white noise. Since many speech sounds are produced by periodic excitation pulses, the sequential adaptive ARMA methods produce poor parameter estimates when such an excitation is present. Miyanaga et al. (1982) introduced a speech analysis algorithm that eliminated the influence of the fundamental frequency of voicing (F_o) using the model reference adaptive system (MRAS). This method used adaptive algorithms to estimate the input and the ARMA parameters simultaneously. This results in increased system complexity for the overall adaptive process. Later, Miyanaga et al. (1986) proposed a model identification system (MIS) that alleviated the influence of the input excitation

on the speech production model. This method also estimated both the ARMA parameters and the input excitation simultaneously by using a recursive algorithm that represented an extended form of the Kalman filter algorithm. There was no weighting control of the error signal in this method because the forgetting factor was unity during the adaptation process, making this method applicable to stationary processes only. However, since the speech signal parameters may change rapidly, as during the transitions between vowels and consonants, a variable forgetting factor can help overcome some of these analysis difficulties.

Ting (Ting, 1989; Ting and Childers, 1990) shows that an adaptive weighted recursive least squares algorithm with a variable forgetting factor (WRLS-VFF) can accurately estimate and track formants and antiformants and their bandwidths. Furthermore, he demonstrated that this technique can also accurately estimate the glottal volume velocity waveform using the formant estimates obtained with the WRLS-VFF algorithm and using the same algorithm to perform inverse filtering. The VFF can be updated recursively at each sample. The input excitation can be determined by using the VFF along with the estimation error of the WRLS algorithm. Two closed phase speech analysis techniques, namely, adaptive formant tracking and glottal inverse filtering, were implemented.

The Ting method (Ting, 1989; Ting and Childers, 1990)used the fixed threshold of the VFF signal for the detection of the pitch period and of the input excitation starting point. Due to the high variation of the VFF signal, the automatic detection of the glottal closed instants and the pitch period could not be performed in Ting's method. Also he used the off-line threshold in the two-pass and two-channel analysis methods for the detection of the glottal closed instants and the pitch period. An addition problem with Ting's method is the inability to automatically estimate the antiformants present in nasals, nasalized vowels, mixed sound, or speech corrupted by noise without classifying the V/U/M/N/S categories in the beginning of the algorithm.

1.4 Research Objectives

The main object of this study is to develop a complete analysis/synthesis system. This system may be applied for voice conversion, synthesis of high quality speech, and visual training aids for the hearing impaired, etc.

In this study, we shall concentrate on the parametric linear source-filter model upon which the formant synthesizer is based. The model falls into two classes:

1) source parameters - fundamental frequency, V/U/M/N/S classification, and glottal waveshape specification parameters

2) tract parameters - formant/anti-formant frequencies, amplitude, and bandwidths.

In this paper we design a more reliable acoustic segmentizer in the fixed frame analysis interval as well as in the pitch synchronous analysis, capable of segmentizing input utterances into the five categories of voiced, unvoiced, mixed, nasal, and silence. The pitch synchronous analysis performed over either one period long frame or closed glottis regions will produce the best result for an analysis/synthesis system and will be used in our system.

New pitch and closed phase instant detection algorithms in the time domain, which provide pitch estimates on a period by period basis and reliable pitch detections in quasi-periodic as well as in aperiodic speech signals, will be described. The pitch smearing effect inherent in speech signal based methods is avoided. The locations of the glottal closing instants allow the isolation of individual periods of the speech waveform from closure to closure. Within each such period, the location of the opening instant is used to further divide the speech signal into closed glottis segments. Such fine segmentation of the speech waveform is essential for the speech analysis/synthesis system.

We also design a closed phase adaptive algorithm for estimating ARMA parameters, input pulse train for voiced speech, and input white noise for the unvoiced

speech at the same time so that formants and antiformants of speech can be correctly estimated. We can obtain correct formants and bandwidths without the influence of pitch, that is, correct parameters of the speech production model. This algorithm should produce an unbiased estimate for stationary signal analysis and track fast parameter changes for nonstationary signals.

The algorithm for conducting the automated glottal inverse filter analysis is studied. This algorithm may accurately estimate the glottal v-v waveform using the formant estimates obtained with the closed phase adaptive algorithm and using the same algorithm to perform inverse filtering. The measuring and modeling voice source characteristics are to model the glottal volume velocity waveform and measure the parameters of the model from the inverse filtered v-v waveform. These parameters are used to produce a modeled glottal v-v waveform for the input of the formant synthesizer and used to classify the different voice types.

For the application, the automatic classification algorithm of different voice types using the parameters of a source model is studied. The LF model parameter sets and a pattern classifier to classify an unknown voice type into a known category of voice type automatically will be used.

1.5 Description of Chapters

Chapter 2 describes the design of this research. It includes the data base used and the techniques associated with reference data collection and processing.

In Chapter 3, we describe the WRLS-VFF-VT algorithm and implementation procedure. While this chapter is based on previous results, we do derive a new procedure for updating the variable forgetting factor and the variable threshold. In addition we derive a new algorithm for simultaneously estimating the WRLS input as well as the model parameters.

In Chapter 4, we present the new pitch detection algorithms which are the EGG based, the LP error based and the VFF based methods. A pathologic speech mode and a normal speech mode will be examed for the pitch detection. These methods provide the pitch period on a period by period basis. Then, the three pattern classification methods for making the V/U/M/N/S decision and the features for use in the five-way classifiers are discussed and described. The pitch synchronous V/U/M/N/S algorithm will also be discussed.

In Chapter 5, we evaluate the performance of the WRLS-VFF-VT algorithm for the formant/antiformant and their bandwidths tracking. We generate several synthetic signals based on both AR and ARMA models and use these signals and real speech signals to test the WRLS-VFF-VT algorithm signals. We then compare our results with other algorithms. We also describe new glottal inverse filtering methods using WRLS-VFF-VT algorithm and compare the algorithm to other methods. Our methods are shown to provide a reliable glottal volume-velocity waveform estimate automatically from the speech signal.

In Chapter 6, our analysis/synthesis system will be tested and evaluated. We synthesize several speech tokens by the formant synthesizer based on the results of our analysis procedure. We also discuss the extraction of the voice source models and their estimation from speech signals. Application of our analysis system in classifying automatically different voice type of speech signals using the VQ classifier is considered.

Chapter 7 gives our conclusions, summary of results, application areas, and recommendations for future study.

In the Appendix, a comparison of results between the ARMA model with WRLS-VFF-VT and the Gamma model with ALMS (Adaptive Least Mean Square) algorithm is given.

CHAPTER 2
RESEARCH DESIGN

2.1 Overview of Research

For speech scientists and engineers, one of the most important objectives of studying human speech production phemonena is to derive appropriate parameters for speech analysis and synthesis models. Speech analysis is concerned with the estimation, from the speech signal, of the parameters of a model for speech production consisting of a slowly time-varying linear system excited by either quasi-periodic (glottal) pulses or random noise. Thus, the basic problems of speech analysis are speech segmentation like V/U/S or V/U/M/S, pitch period estimation for voiced speech, estimation of the glottal pulse shape, and estimation of vocal tract transmission properties.

Figure 2−1 depicts a block diagram of the overall analysis for synthesis scheme using the WRLS-VFF-VT algorithm. It is characterized by selecting the input data as one-channel (only speech signal) or as two-channels (speech and EGG signals). Depending on the input signal type, the pitch detection method, we label these analysis systems by three different names: EGG based, LP error based, and VFF based WRLS-VFF-VT methods.

For the EGG based WRLS-VFF-VT, EGG and DEGG (differentiated EGG) can be used to isolate individual pitch periods and to determine glottal closed intervals using the EGG- based pitch detection method discussed in section 4.2.

For the LP error based WRLS-VFF-VT, the location of the main pulses of the LP error signal derived in the LPC analysis and the inverse filtering can be used to

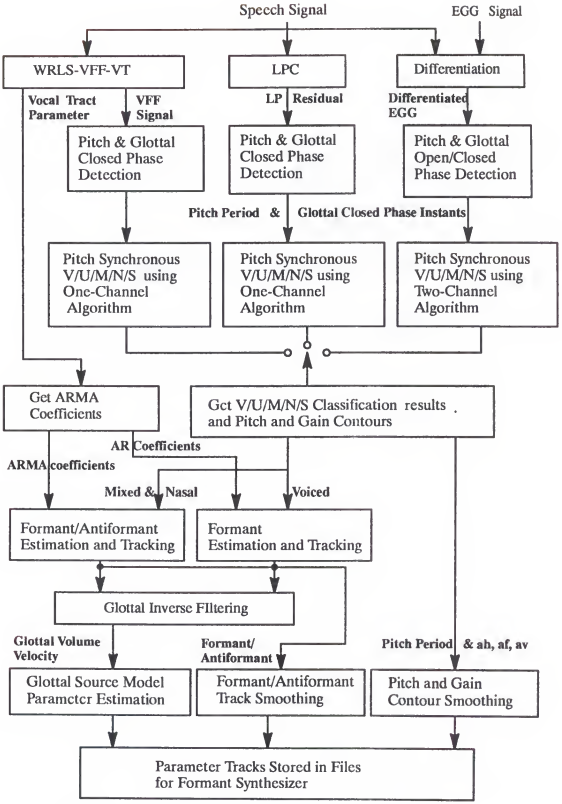


Figure 2−1. Overview of analysis for synthesis algorithm.

detect individual pitch periods and the glottal closure intervals using the LP error-based method discussed in section 4.2.

For the VFF based WRLS-VFF-VT which does not require the two channels of data or the two-pass of the data, the VFF signal derived in the WRLS-VFF-VT algorithm can be used directly to decide the pitch periods and the glottal closed intervals using the VFF-based pitch method discussed in section 4.2.

After the pitch determination procedure, the pitch-synchronous V/U/M/N/S classification described in section 4.3 follows. Using the glottal closed phase intervals, the vocal tract parameters can be extracted from the output of the WRLS-VFF-VT algorithm.

The selection of the AR and ARMA model for the WRLS-VFF-VT algorithm is dependent on the results of V/U/M/N/S classification. If an analysis frame involves the mixed and the nasal sounds, an ARMA model is selected. Otherwise, an AR model is selected.

The formant/anti-formant resonances of the vocal tract are estimated by solving the roots of the LP polynomial, and then shaping the formant/anti-formant structure by empirical rules. These rules include: 1) discarding the roots with center frequencies under 250 Hz, 2) discarding the roots with a Q less than one, and 3) merging two adjacent roots.

The refined formant/antiformant resonances are then used to construct the vocal tract transfer function, which are used in the glottal inverse filtering (GIF) procedure. The direct output of the GIF operation is the differential glottal volume-velocity waveform.

Sometimes, a smoothing procedure is needed because the formant and the pitch contours have spurious or missing peaks.

Within an analysis frame (corresponding to one pitch period) of the inverse filtered differentiated glottal flow waveforms, the LF model parameters can be

extracted. The timing parameters of the LF model are closely related to the glottal waveshape factors.

For the formant synthesizer, all necessary parameters required are the following: 1) the formant/anti-formant frequencies, bandwidths, and amplitude 2) the gain contours for voiced (av), fricative (af) and plosive (ah) excitation source. 3) the glottal model parameters to generate a glottal flow signal.

This integrated analysis package automatically processes through the pitch detection, the pitch-synchronous V/U/M/N/S classification, the formant/antiformant and their bandwidths tracking using WRLS-VFF-VT algorithm and the glottal inverse filtering for the LF model parameters. These analysis results are used in the inputs of the formant synthesizer.

2.2 Experimental Data Base

2.2.1 Subjects and Tasks for Analysis

The data base used in this research consisted of recordings of two vowels and sentences from several speakers of different voice types. Each speech data record was previously categorized by professional speech scientists into one of three different voice types including breathy, vocal fry, and modal.

Three categories of subjects served in this research: (1) normal subjects (CKL, DRW) who had no history of vocal disorders or laryngeal pathology, (2) patients with vocal disorders (EDR, JMS, JTO) whose voices were evaluated by experienced speech pathologists, (3) experienced speech pathologists (DMH, GPM). All subjects were male. Female subjects were excluded to avoid a gender factor in the research.

The experimental tasks for each subject were

(1) sustained vowels /i/ and /a/ using an Electro-Voice RE-10 microphone, and a Bruel & Kjaer model 4133 condenser microphone,

(2) counting from one to ten with comfortable pitch and loudness,

(3) counting from one to ten with progressive increase in loudness,

(4) singing the chromatic scale using "la",

(5) three sentences ("We were away a year ago", "Early one morning a man and a woman ambled along a one mile lane", and "Should we chase those cowboys?"). Tasks two through five used the Electro-Voice RE-10 microphone. Two speech pathologists mimicked various voice types (modal, breathiness, vocal fry, hoarseness) for the same tasks. The normal subject CKL also mimicked vocal fry to perform task one. Table 2−1 lists the data base, that were analyzed for this study, according to the subjects and the tasks they performed.

2.2.2 Data Base for the Application of Analysis/Synthesis

To create the data base on the application of the analysis/synthesis system in this study, five sentences were selected based on their phonetic contents. These five sentences are following:

Sentence 1: We were away a year ago.

Sentence 2: Early one morning a man and a woman ambled along a one mile lane.

Sentence 3: Should we chase those cowboys?

Sentence 4: That zany van is azure.

Sentence 5: We saw the ten pink fish.

Three male and three female normal speakers were asked to utter these five sentences one by one with comfortable speed, tone, and loudness using the Electro-Voice RE-10 microphone in an IAC (Industrial Acoustics Company) sound

booth. With six speakers and five sentences, the total number of sentences generated for the study was thirty. In terms of phonetics, Sentence 1 is composed of all voiced, vocalic sounds. Sentence 2 adds nasals and liquids. Sentence 3 adds fricatives and affricates, while Sentence 4 contains all the voiced fricatives of English. Finally, Sentence 5 has unvoiced fricatives and plosives. All the file names and their lengths, as stored in our computer system, are shown in Table 2−2. (File names for EGG data have the same names as corresponding speech data except that they have extensions beginning with 'e' instead of 's'.) For convenience, the sentences will be referred to with names like 'sentence 1-a' instead of by file names, such as 'nraaan025.smst'.

2.2.3 Data Collection

All data recordings were done inside an Industrial Acoustics Company (IAC) single-wall sound room. The speech signals collected have a high signal-to-noise ratio (SNR) and noise can be effectively ignored. The speech and electroglottographic (EGG) signals were collected simultaneously. A microphone (an Electro-Voice RE-10 dynamic cardioid microphone or a Bruel & Kjaer (B&K) model 4113 condenser microphone, depending on the task recorded) was located at a fixed distance of 6 inches from the speaker's lips. The electroglottograph used was a Synchrovoice Inc. model.

Before digitization, the speech and EGG signals were bandlimited to 5 kHz by anti-aliasing, passive, elliptic filters with a minimum stopband attenuation of -55 dB and a passband ripple of ±0.2 dB. Both signals were then amplified by a Digital Sound Corporation DSC-240 audio control console. The synchronized speech and EGG signals were directly digitized at a sampling frequency of 10 kHz per channel by a Digital Sound Corporation DSC-200 A/D and D/A system with 16-bit precision.

Table 2−1. The data base for speech analysis

Subject	Sex/Age	Phonation Type	Data File	Contents	Microphone
DMH	M/37	modal voice	dmhn003c	/i/	E
			dmhn012c	/a/	E
			dmhn025c	S1	E
			dmhn028	/a/	B
DRW	M/23	modal voice	drwn003c	/i/	E
			drwn012c	/a/	E
			drwn025c	S1	E
			drwn028	/a/	B
CKL	M/31	modal voice	modb1	/i/	B
			modb2	/a/	B
			mode1c	/i/	E
			mode2c	/a/	E
DMH	M/37	mimicked breathy voice	dmhp010c	/i/	E
			dmhp012c	S1	E
EDR	M/22	pathological breathy voice	edrp003c	/i/	E
			edrp005c	S1	E
			edrp010	/i/	B
GPM	M/79	mimicked breathy voice	gpmp003c	/i/	E
			gpmp005c	S1	E
JMS	M/30	pathological breathy voice	jmsp003c	/i/	E
			jmsp005c	S1	E
			jmsp010	/i/	B
CKL	M/31	mimicked fry voice	fryb1	/i/	B
			fryb2	/a/	B
			frye1c	/i/	E
			frye2c	/a/	E
JTO	M/21	pathological fry voice	jtop003c	/i/	E
			jtop005c	S1	E
			jtop010	/i/	B

*Contents: S1 = "We were away a year ago."
*Microphone Type: E = Electro-Voice RE-10, B = B&K4113

Table 2−2. The data base for speech analysis

Sentence	(Sex)	File Name	Data Length
Sentence 1-a	(M)	nraaan025.smst	18688
Sentence 1-b	(M)	nrdrwn025.smst	18942
Sentence 1-c	(M)	nrjrsn025.smst	20223
Sentence 1-d	(F)	nrcxon025.sfst	19968
Sentence 1-e	(F)	nrbemn025.sfst	19968
Sentence 1-f	(F)	nrmbkn025.sfst	21760
Sentence 2-a	(M)	nraaan026.smst	45056
Sentence 2-b	(M)	nrdrwn026.smst	43520
Sentence 2-c	(M)	nrjrsn026.smst	42496
Sentence 2-d	(F)	nrcxon026.sfst	38656
Sentence 2-e	(F)	nrbemn026.sfst	41472
Sentence 2-f	(F)	nrmbkn026.sfst	43264
Sentence 3-a	(M)	nraaan027.smst	18432
Sentence 3-b	(M)	nrdrwn027.smst	20992
Sentence 3-c	(M)	nrjrsn027.smst	20224
Sentence 3-d	(F)	nrcxon027.sfst	19968
Sentence 3-e	(F)	nrbemn027.sfst	20224
Sentence 3-f	(F)	nrmbkn027.sfst	20224
Sentence 4-a	(M)	nra1an001.smwt	21759
Sentence 4-b	(M)	nrd1cn001.smwt	22528
Sentence 4-c	(M)	nrd1hn001.smwt	21504
Sentence 4-d	(F)	nrd1hn001.sfwt	26624
Sentence 4-e	(F)	nrm1kn001.sfwt	20480
Sentence 4-f	(F)	nrn1sn001.sfwt	21504
Sentence 5-a	(M)	nra2an001.smwt	21504
Sentence 5-b	(M)	nrd2cn001.smwt	24576
Sentence 5-c	(M)	nrm1gn001.smwt	20992
Sentence 5-d	(F)	nrd2hn001.sfwt	20480
Sentence 5-e	(F)	nrm2kn001.sfwt	25600
Sentence 5-f	(F)	nrb1cn001.sfwt	20736

2.2.4 Microphone Characteristics

When the speech signal is used for glottal source estimation, the recording device should have a good low-frequency response. The reason for this is that the glottal source waveform, which is to be estimated, has its major energy components at low frequencies (dc to 1 kHz). Of the two microphones used to measure the sound pressure waveforms, the B&K 4133 condenser microphone has the best low-frequency response. Its amplitude response is within ±1 dB down to 20 Hz, and its phase response is linear. The -3 dB low-frequency cut-off is approximately 10 Hz. Because of this good low-frequency characteristic, the B&K 4133 condenser microphone is also sensitive to low-frequency breath and ambient noise, which may cause problems in speech analysis. Therefore, the Electro-Voice RE-10 microphone was used to collect most of the speech data.

The Electro-Voice RE-10 microphone has a good frequency response at frequencies above 50 Hz, but attenuates the low-frequency components below 50 Hz. When compared to the B&K 4133 condenser microphone, the obvious drawback of the Electro-Voice RE-10 microphone is the lack of good low-frequency response. Thus, speech data collected by using the Electro-Voice RE-10 microphone had to be corrected to compensate for the low-frequency distortion, based upon the characteristics of the B&K 4133 condenser microphone (Wong, 1991). In this research we used only corrected speech data (if recorded through the Electro-Voice RE-10 microphone) as well as data obtained by using the B&K 4133 condenser microphone.

2.3 Preprocessing of Data

2.3.1 Demultiplexing and Trimming the Data

The collected data are two-channel (speech and EGG) multiplexed signals sampled at 20 kHz, and contain a large portion of silence at the beginning and end of each file. The signal is demultiplexed to produce both the digitized speech and EGG signals with the sampling frequency of 10 kHz. These demultiplexed signals are trimmed to get rid of unnecessary surplus silence data at the beginning and end of each utterance to save computer memory. While trimming, the operator left at least five silence frames at the beginning of each utterance. These frames would be used to obtain the statistics for silence such as the average zero crossing rate and the average energy level, which are essential to the five-way classification algorithm.

2.3.2 Synchronization of Data

The microphone was kept 6 inches (15.24 centimeters) away from the speaker's lips to reduce breath noises and to simplify the alignment procedure. Synchronization of the speech and EGG waveforms is necessary to account for the time delay while the speech signal travels from the vocal folds to the microphone. This time delay can be expressed as follows.

$$T_d = (VT_l / C_{VT}) + (SM_l / C_{air}) \qquad (2.1)$$

where T_d is the time delay in seconds and VT_l is the vocal tract length in centimeters. The distance from the speaker's lips to the microphone, 15.24 centimeters in this study, is denoted as SM_l. C_{VT} and C_{air} are for the velocities of sound in the vocal tract and in air, respectively. If we select typical values of these parameters, e.g., VT_l

of 17.0 centimeters (for adult male subjects), C_{VT} of 35300 cm/sec (Flanagan, 1972; Dunn, 1950), and C_{air} of 34400 cm/sec, the T_d obtained is 0.925 milliseconds. Hence the number of data points to be discarded from the beginning of the speech record is nine.

The matter of variation in vocal tract lengths among adult males was largely resolved with the 17.0 centimeter compromise. Equation (2.1) shows that a nine-data-point correction is actually appropriate for vocal tract lengths from 14.4 to 17.9 centimeters long. On the other hand, the average length of the vocal tracts among adult females is known to be 14.0 centimeters (Borden and Harris, 1984), and this leads to a one-data-point misalignment of the speech and EGG signals. This misalignment does not cause any serious problem in the design of a reliable four-way classification algorithm because a segment size of 100 data points would be used. Examination of the data also supported the use of this nine-data-point correction for adult speakers.

2.4 Electroglottograph (EGG)

The ElectroGlottoGraph (EGG) is an instrument designed to register the vocal fold vibration as a time-varying signal. The EGG measures the radio-frequency (RF) impedance across the larynx, and hence the amplitude variations of the EGG signal are generally thought to be representative of the area of contact of the vocal folds. An objective of this device is to provide a measure of vocal fold activity de-coupled from the effects of the supra-glottal system. A comprehensive review about the EGG instrumentation with the waveform interpretation and the application was given by Krishnamurthy and Childers (1986).

A pair of electrodes is applied to the neck at the level of the larynx. A high frequency (about 5 MHz) current passes from one electrode through the neck and is

picked up by the other electrode. As the subject phonates, the opening and closing of the vocal folds changes the electrical impedance of the neck in the region of the electrodes. This modulates the RF current, which is then demodulated using a detector to yield the EGG signal.

The EGG indicates the electrical impedance through the neck at the level of the larynx and thus monitors variations in vocal fold contact - glottal closure is associated with a reduction in tissue impedance. A time lag for the acoustic propagation delay from the glottis to the microphone is applied to the EGG signal when it is compared with the speech signal. The magnitude of the EGG signal is inversely proportional to vocal fold contact area, so that an increase in amplitude denotes glottal opening. The steep negative slope of the EGG signal associated with glottal closure occurs over only one or two sample points. Glottal opening occurs more slowly and makes the true opening point more difficult to determine accurately.

The EGG does not reflect a direct measure of glottal area. It is postulated that the tissue impedance is inversely proportional to the lateral contact area of the vocal folds. The EGG channel can (1) help solve the deconvolution problem of inverse filtering the speech signal, (2) improve voiced, unvoiced, and silence detection and fundamental frequency estimation, and (3) facilitate spectral estimation and formant tracking (Krishnamurthy and Childers, 1986). The EGG-based pitch detection scheme provides the pitch on a period-by-period basis. The two-channel (speech and EGG) speech analysis technique provides computational and performance improvements over "speech only" analysis methods because of the added EGG channel.

CHAPTER 3
THE WEIGHTED RECURSIVE LEAST SQUARES ALGORITHM WITH A VARIABLE FORGETTING FACTOR AND A VARIABLE THRESHOLD (WRLS-VFF-VT)

3.1 Introduction

The successful application of system identification techniques in adaptive control motivated us to apply the same techniques to signal processing. Of particular interest is the fact that some commonly used parameter estimation algorithms such as recursive maximum likelihood, recursive least squares and least squares lattice, are capable of estimating ARMA parameters, and not just AR parameters.

According to Landau (1976), recursive identification of the dynamic parameters of a process, as well as tracking the process parameters when they are time-varying, can be formulated as a model reference adaptive problem. The process (unknown system) to be identified represents the reference model. The adjustable system is constituted by an adjustable model (also called an estimation model) having the same structure as the mathematical model of the process whose parameters are controlled by an adaptive mechanism (which implements an identification algorithm).

The application of the adaptive system identification algorithm to the ARMA parameter estimation needs some assumptions.

1. Assume the unknown system to be an ARMA process driven by an inaccessible white noise process. This also leads to an adaptive infinite impulse response (IIR) filter.

2. Since the input to the estimation model is not available at each step, an input estimation algorithm is required to estimate the input before executing the adaptive process.
3. The variable forgetting factor (VFF) and the variable threshold (VT) proposed in this study and the prediction error of the adaptive process can be used to estimate the input excitation of the unknown system.

Using the weighted recursive least squares (WRLS) algorithm for the ARMA parameter estimation is proposed for this study. Its block diagram is shown in Figure 3−1.

3.2 Algorithm Description

3.2.1 Background for the WRLS Algorithm

We will assume that the speech signal is generated by an ARMA model represented by the following:

$$y_k = -\sum_{i=1}^{p} a_i(k)\, y_{k-i} + \sum_{j=1}^{q} b_j(k) u_{k-j} + u_k \qquad (3-1)$$

where y_k denotes the k-th sample of the speech signal, u_k is the input excitation to the ARMA model, (p, q) are the order of the poles and zeros, respectively, of the ARMA model, and $a_i(k)$ and $b_j(k)$ are the time-varying AR and MA parameters, respectively. Measurement noise is ignored in this model but could be included (Miyanaga et al., 1982, 1986). Model order selection techniques have been proposed by numerous authors (Akaike, 1974; Kay, 1987; Marple, 1987; Rissanen, 1978). Here, we assume that the values of p and q can be predetermined. Note that the measured speech signal, y_k, depends on the input, u_k. The excitation, u_k, is usually considered to be white

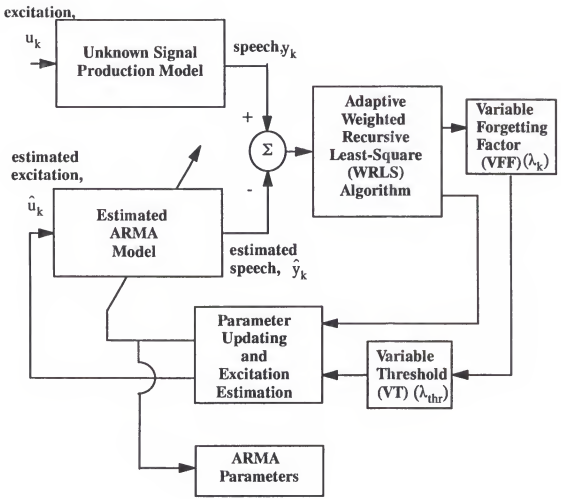


Figure 3−1. Block diagram of the WRLS-VFF-VT algorithm for the ARMA parameter estimation.

Gaussian noise. In this paper we allow u_k to be either a zero-mean, white, Gaussian noise process, u_k^w, with variance $\sigma^2_{u,k}$, or a train of periodic pulses, u_k^p. We develop an extension to an algorithm that estimates both the ARMA parameters, type of input excitation, and the VFF. We must estimate the input excitation, u_k, so that the ARMA parameters can be estimated accurately from y_k. An estimation method for u_k based on the variable forgetting factor of the WRLS algorithm will be given later. For the present we assume that an estimate for u_k, denoted as $\hat{u}_k$, is available.

Let us define a parameter vector, θ_k, its estimate, $\hat{\theta}_k$, and a data vector, ϕ_k, and its estimate, $\hat{\phi}_k$, by the following equations:

$$\theta_k{}^t = [a_1(k),\ldots,a_p(k),b_1(k),\ldots,b_q(k)] \qquad (3-2)$$

$$\hat{\theta}_k{}^t = [\,\hat{a}_1(k),\ldots,\hat{a}_p(k),\hat{b}_1(k),\ldots,\hat{b}_q(k)] \qquad (3-3)$$

$$\phi_k{}^t = [-y_{k-1},\ldots,-y_{k-p},u_{k-1},\ldots,u_{k-q}] \qquad (3-4)$$

$$\hat{\phi}_k{}^t = [-y_{k-1},\ldots,-y_{k-p},\hat{u}_{k-1},\ldots,\hat{u}_{k-q}] \qquad (3-5)$$

where the superscript t denotes transpose, and $\hat{a}_i$ and $\hat{b}_i$ are the estimated ARMA parameters, respectively. Using $(3-2)$ - $(3-5)$ the speech signal, y_k, and its estimate, $\hat{y}_k$, may be expressed as

$$y_k = \phi_k{}^t\theta_k + u_k \qquad (3-6)$$

$$\hat{y}_k = \hat{\phi}_k{}^t\hat{\theta}_k + \hat{u}_k \qquad (3-7)$$

Let r_k be the residual error of the ARMA process, namely,

$$r_k = y_k - \hat{y}_k = y_k - \hat{\phi}_k{}^t \hat{\theta}_k - \hat{u}_k \qquad (3-8)$$

The predicted signal, $\hat{y}_{k/k-1}$, determined from the estimated ARMA parameters and estimated at (k-1) is

$$\hat{y}_{k/k-1} = \hat{\phi}_k{}^t \hat{\theta}_{k-1} \qquad (3-9)$$

Consequently, the prediction error is

$$e_k = y_k - \hat{y}_{k/k-1} - \hat{u}_k \qquad (3-10)$$

(Note that $\hat{u}_k$ is usually assumed to be not available at (k-1) and is set to zero (Miyanaga et al., 1982; 1986). We will address this issue again later and modify the algorithm accordingly.)

The weighted estimation (or residual) error is (Miyanaga et al., 1982; Soderstrom, 1989)

$$\begin{aligned} E_k &= \sum_{i=1}^{k} w(i,k)(y_i - \hat{y}_i)^2 + w(1,k)[\hat{\theta}_k{}^t P_1{}^{-1} \hat{\theta}_k] \\ &= \sum_{i=1}^{k} w(i,k) r_i{}^2 + w(1,k)[\hat{\theta}_k{}^t P_1{}^{-1} \hat{\theta}_k] \end{aligned} \qquad (3-11)$$

where P_1 is an arbitrary real symmetric positive definite matrix. The weighting coefficient w(i, k) is (Miyanaga et al., 1982; Soderstrom, 1989)

$$\begin{aligned} w(i,k) &= \prod_{j=i+1}^{k} \lambda_j \qquad ,i = 1,2,...,k-1 \\ &= 1 \qquad\qquad\quad ,i = k,... \end{aligned} \qquad (3-12)$$

(Note that we have changed the indexing on w(i,k) in (Miyanaga et al., 1982) so that

it agrees with that in (Soderstrom, 1989)) The coefficient λ_j decreases the weight of past estimation errors provided $0<\lambda_j<1$. Note that for fixed $\lambda_j = \lambda$ that w(i,k) becomes an exponentially weighted coefficient, e.g., λ^{k-1}, λ^{k-2}, ..., λ , 1. Consequently, the estimation error, E_k, becomes the exponentially weighted sum of squares of the estimation errors (Cowan, 1985; Friedlander, 1982; Soderstrom and Stoica, 1989), i.e.,

$$E_k = \sum_{i=1}^{k} \lambda^{k-i}(y_i - \hat{y}_i)^2 + \lambda^{k-1} [\hat{0}_k{}^t \, P_1{}^{-1} \, \hat{0}_k] \qquad (3-13)$$

where the last term in (3 − 13) is not considered in (Cowan, 1985; Friedlander, 1982).

Minimizing the least square weighted estimation error, E_k, with respect to the ARMA parameter vector, $\hat{0}_k$, assuming that $\hat{u}_k$ is available, gives (Miyanaga et al., 1982, 1986; Soderstrom, 1989)

Residual error: $$r_k = y_k - \hat{\phi}_k{}^t \, \hat{0}_k - \hat{u}_k \qquad (3-14)$$

Prediction error: $$e_k = y_k - \hat{\phi}_k{}^t \, \hat{0}_{k-1} - \hat{u}_k \qquad (3-15)$$

Gain update: $$K_k = P_{k-1} \, \hat{\phi}_k \, [\lambda_{k-1} + \hat{\phi}_k{}^t \, P_{k-1} \, \hat{\phi}_k]^{-1} \qquad (3-16)$$

Parameter update: $$\hat{0}_k = \hat{0}_{k-1} + K_k \, e_k \qquad (3-17)$$

Covariance matrix: $$P_k = \lambda_k{}^{-1} \, [P_{k-1} - K_k \, \hat{\phi}_k{}^t \, P_{k-1}] \qquad (3-18)$$

The above algorithm updates the ARMA parameters at each instant k. The algorithm has been shown to be stable and to provide a unique solution (Cowan, 1985;

Friedlander, 1982; Miyanaga et al., 1982, 1986; Soderstrom and Stoica, 1989). As analyzed in (Soderstrom and Stoica, 1989), λ must equal unity to obtain convergence. When $\lambda < 1$ the algorithm is more sensitive and parameter estimates change more quickly. Several investigators have let λ vary with time (or index k). One variation selected is to let $\lambda(k)$ tend exponentially to unity,

$$\lambda(k) = 1 - \lambda_0{}^k \, [1 - \lambda(0)] \qquad (3-19)$$

or in recursion form is

$$\lambda(k) = \lambda_0 \, \lambda(k-1) + (1 - \lambda_0) \qquad (3-20)$$

where typical values for λ_0 and $\lambda(0)$ are 0.99 and 0.95, respectively (Friedlander, 1982; Soderstrom and Stoica, 1989). However, we know of no work that has adaptively estimated λ_∞. We do so below.

3.2.2 Adaptive WRLS-VFF-VT Algorithm

For a locally stationary speech production process, the residual error, r_k, in $(3-8)$ or $(3-14)$ will indicate the state of the estimator at each instant k. If the forgetting factor, λ_k, is fixed, i.e., $\lambda_k = \lambda$, and the error is small, then λ should be near unity, allowing the adaptive algorithm to use most of the previous information in the signal. This yields accurate estimates of the various parameters. If, on the other hand, the error is large, then a small λ_k will decrease the weighting of the error, thereby shortening the effective memory length of the estimation process. This allows the parameters to be adjusted with the most recent data, and will reduce the error. A discussion of this situation appears elsewhere (Soderstrom and Stoica, 1989).

As discussed above, previous work has fixed λ_k or let λ_k vary exponentially. We derive a procedure to compute and update λ_k. We begin with the procedure developed by Miyanaga et al. (1982), who defined the estimation error as in Eqs. (3 − 11) and (3 − 12). Note that $w(i,k)=\lambda_{i+1} \ldots \lambda_k$, $w(k-1,k)=\lambda_k$, $w(k,k)=1$, $w(n,k)=1$ for $n \geq k$, and $w(1,k-1)=\lambda_2\lambda_3 \ldots \lambda_{k-1}$. Thus, Eq. (3 − 11) becomes

$$E_k = \sum_{i=1}^{k} w(i,k)\ (y_i - \hat{y}_i)^2 + w(1,k)[\hat{\theta}_k{}^t\ P_1{}^{-1}\ \hat{\theta}_k]$$

$$= \sum_{i=1}^{k-1} w(i,k)\ (y_i - \hat{y}_i)^2 + (y_k - \hat{y}_k)^2 + w(1,k)[\hat{\theta}_k{}^t\ P_1{}^{-1}\ \hat{\theta}_k] \qquad (3-21)$$

Similarly,

$$E_{k-1} = \sum_{i=1}^{k-1} w(i,k-1)\ (y_i - \hat{y}_i)^2 + w(1,k-1)[\hat{\theta}_{k-1}{}^t\ P_1{}^{-1}\ \hat{\theta}_{k-1}] \qquad (3-22)$$

Thus, we may express E_k recursively as

$$E_k = \lambda_k E_{k-1} + (y_k - \hat{y}_k)^2 + w(1,k)[\hat{\theta}_k{}^t\ P_1{}^{-1}\ \hat{\theta}_k] - \lambda_k w(1,k-1)[\hat{\theta}_{k-1}{}^t\ P_1^{-1}\hat{\theta}_{k-1}]$$

$$= \lambda_k E_{k-1} + (y_k - \hat{y}_k)^2 + w(1,k)[\hat{\theta}_k{}^t\ P_1{}^{-1}\ \hat{\theta}_k - \hat{\theta}_{k-1}{}^t\ P_1^{-1}\hat{\theta}_{k-1}] \quad (3-23)$$

Then re-arranging, we have

$$\lambda_k = \frac{E_k}{E_{k-1}} - \frac{(y_k - \hat{y}_k)^2}{E_{k-1}} - \frac{w(1,k)}{E_{k-1}}[\hat{\theta}_k{}^t\ P_1{}^{-1}\ \hat{\theta}_k - \hat{\theta}_{k-1}{}^t\ P_1^{-1}\hat{\theta}_{k-1}] \quad (3-24)$$

The numerator of the second term on the right side of (3 − 24) can be expressed as follows

$$y_k = c_k + \hat{y}_{k/k-1} + \hat{u}_k = c_k + \hat{\phi}_k{}^t\, \hat{\theta}_{k-1} + \hat{u}_k \qquad (3-25)$$

$$\hat{y}_k = \hat{\phi}_k{}^t\, \hat{\theta}_k + \hat{u}_k \qquad (3-26)$$

$$y_k - \hat{y}_k = e_k + \hat{\phi}_k{}^t\, [\hat{\theta}_{k-1} - \hat{\theta}_k] \qquad (3-27)$$

Since

$$\hat{\theta}_k = \hat{\theta}_{k-1} + K_k e_k \qquad (3-28)$$

then

$$y_k - \hat{y}_k = e_k\, (1 - \hat{\phi}_k{}^t\, K_k) \qquad (3-29)$$

Then we finally have the following approximation for computing and updating λ_k as

$$\lambda_k = \frac{E_k}{E_{k-1}} - \frac{e_k{}^2}{E_{k-1}}[1 - \hat{\phi}_k{}^t\, K_k]^2 \qquad (3-30)$$

where we have assumed the third term in $(3-24)$ becomes negligible with increasing k since $w(1,k)=\lambda_2\lambda_3 \ldots \lambda_k \ll 1$ for $0 \leq \lambda_i < 1$. Furthermore, the difference in the third term will be zero in the algorithm we develop below for simultaneously estimating the pulse input and the model parameters. When the input is noise this term should be small for increasing k since λ_k will approach unity and the estimate for $\hat{\theta}_k$ will stabilize. A difficulty with calculating $(3-30)$ is that we need E_k and E_{k-1} before we can estimate λ_k. A strategy for calculating λ_k may be defined by requiring E_k to be constant such that

$$E_k = E_{k-1} = \ldots = E_1. \qquad (3-31)$$

In other words, the forgetting factor will compensate at each step k for the new error

information in the latest measurement, thereby insuring that the estimation is always based on the same error information. Thus from (3 − 30) and (3 − 31) we have

$$\lambda_k = 1 - \frac{e_k^{\,2}}{E_1}[1 - \hat{\phi}_k^{\,t}\, K_k]^2 \qquad (3-32)$$

Consequently, the WRLS-VFF-VT algorithm can be specified by a set of equations similar to those for the WRLS algorithm in (3 − 14) - (3 − 18), but with the constant weighting factor, λ_k, estimated by (3 − 32). The effective memory of the algorithm can be defined as (Cowan and Grant, 1985)

$$N = 1/(1 - \lambda_k) \qquad (3-33)$$

Other definitions include $N = \lambda_k/(1-\lambda_k)$ (Soderstrom and Stoica, 1989).

If λ_k becomes small, then the memory also becomes small. In some applications we may require a pre-determined memory size. For such cases, we recommend a minimal λ_k be defined as

$$\lambda_{min} = 1 - \frac{1}{N_a}\ , \text{if } \lambda_k < \lambda_{min}\ , \text{then } \lambda_k = \lambda_{min} \qquad (3-34)$$

where $N_a = p+q$ is the total number of the filter coefficients in the ARMA model. Otherwise λ_k can be calculated from (3 − 32) by calculating E_1 over one or two analysis frames using a block data method. Then with an estimate for E_1 available, the other terms in (3 − 32) can be found, giving λ_k. In other words, we may decide to use some minimal value for k, instead of k=1, to obtain a more realistic estimate for the initial estimate of E_k, the weighted sum of the residual errors. We might do this initial estimation using an initial block of data and then begin the recursion process.

3.2.3 WRLS-VFF-VT Algorithm with Input Estimation

The input excitation u_k to a speech production process can be either pulse trains for voiced sounds or white noise for fricatives. We express the input sequence as follows:

$$u_k = u_k{}^p \qquad (3-35)$$

or

$$u_k = u_k{}^w \qquad (3-36)$$

where $u_k{}^p$ represents the pulse input and $u_k{}^w$ is the white noise input.

Several methods have been proposed to estimate the input u_k in the recursive ARMA parameter estimation algorithm. Morikawa and Fujisaki (1982) and Friedlander (1982) used the estimated residual error r_k at the instant k as the estimated input u_k, namely

$$\hat{u}_k{}^w = r_k = y_k - \hat{y}_k \qquad (3-37)$$

This method is based on the fact that the driving source to the ARMA process is a pure white noise. Miyanaga et al. (1982) used the ratio of the variance of zero-mean white noise at k and k+1 to estimate the forgetting factor (FF). For speech analysis applications we need to estimate the presence of either $\hat{u}_k{}^p$ or $\hat{u}_k{}^w$.

The proposed input estimation method uses the FF as a reference to examine the input condition. This can reduce the algorithm complexity since only one adaptive algorithm is used instead of two as in (Miyanaga et al., 1982). Moreover, the FF can be obtained from the adaptive process, and the complexity of the algorithm will be shown to be no greater than that of other similar algorithms. However, the residual and prediction error estimates must be changed from those used in Sections IIA and IIB

where we assumed that an estimate for u_k was available. Here we must estimate the residual and predictions errors as well as the gain, ARMA parameters, the covariance matrix, and the excitation, u_k. Furthermore, we must decide whether $\hat{u}_k$ is $\hat{u}_k{}^p$ or $\hat{u}_k{}^w$.

First, we modify our definitions of the residual and prediction errors to account for the fact an estimate for the excitation, $\hat{u}_k$, is not available at k. Thus, we have

$$r_k = y_k - \hat{y}_k = y_k - \hat{\phi}_k{}^t \hat{\theta}_k \qquad (3-38)$$

and

$$e_k = y_k - \hat{y}_{k/k-1} = y_k - \hat{\phi}_k{}^t \hat{\theta}_{k-1} \qquad (3-39)$$

The parameter update estimate remains the same as before, namely

$$\hat{\theta}_k = \hat{\theta}_{k-1} + K_k e_k \qquad (3-40)$$

as does the residual error and the gain

$$r_k = y_k - \hat{y}_k = e_k (1 - \hat{\phi}_k{}^t K_k) \qquad (3-41)$$

$$K_k = P_{k-1}\hat{\phi}_k [\lambda_{k-1} + \hat{\phi}_k{}^t P_{k-1}\hat{\phi}_k]^{-1} \qquad (3-42)$$

We then update the estimate for λ_k with

$$\lambda_k = 1 - \frac{e_k{}^2}{E_1} [1 - \hat{\phi}_k{}^t K_k]^2 \qquad (3-43)$$

The covariance matrix is updated as follows

$$P_k = \lambda_k{}^{-1} [P_{k-1} - K_k \hat{\phi}_k{}^t P_{k-1}] \qquad (3-44)$$

where in the equations (3 – 40) - (3 – 44) we use the prediction error specified in (3 – 39).

From (3 − 43), we see that an increase in the prediction error results in a decrease in λ_k. A small value of λ_k indicates that the input has an abrupt change (e.g., pulses). Hence, we can determine the time of occurrence of a pulse by determining the instant at which λ_k falls below a threshold value λ_{thr}. A strategy for choosing the threshold value λ_{thr} may now be defined by letting

$$E_k = 1/M \sum_{i=0}^{M-1} \lambda_{k-i} \qquad (3-45)$$

$$\text{If } E_k < 0.9, \text{ then } \lambda_{thr} = 0.99 * E_k \qquad (3-46)$$

$$\text{If } E_k > 0.9, \text{ then } \lambda_{thr} = 0.9 * E_k \qquad (3-47)$$

$$\text{If } \lambda_{thr} > \lambda_{min}, \text{ then } \lambda_{thr} = \lambda_{min} \qquad (3-48)$$

The magnitude of the pulse can be approximately given by the prediction error e_k at the estimated time of the input pulse (Miyanaga et al., 1982), i.e, at the instant k, where $\hat{u}_k = \hat{u}_k{}^p$ and $\hat{u}_k{}^w = 0$ and the prediction error should be the estimated input excitation for a pulse, yielding

$$\hat{u}_k{}^p = y_k - \hat{\phi}_k{}^t \hat{\theta}_{k-1} \qquad (3-49)$$

For white noise input, λ_k is close to unity upon convergence (Soderstrom and Stoica, 1989). Under this condition the residual error r_k of the adaptive process can be used as the estimate of the white noise input, $\hat{u}_k{}^w$, as indicated in Morikawa's method (Morikawa and Fujisaki, 1982, 1984), i.e., from (3 − 41)

$$r_k = y_k - \hat{y}_k = e_k (1 - \hat{\phi}_k{}^t K_k) = \hat{u}_k{}^w \qquad (3-50)$$

where $\hat{u}_k{}^p = 0$. For speech analysis the WRLS-VFF-VT algorithm with input estimation is summarized in Table 3−1. The data/excitation vector, $\hat{\phi}_k{}^t$, is updated

for the next estimate at (k+1) using the new data and the estimate in Table 3−1. The algorithm in Table 3−1 differs from previous algorithms because we 1) update VFF, λ_k, and VT, λ_{thr} 2) let the prediction error, e_k, be the estimate for the pulse magnitude, $\hat{u}_k{}^p$, and 3) let the residual error, r_k, be the estimate for the noise excitation, $\hat{u}_k{}^w$. The algorithm does not need a separate voiced/unvoiced decision step for it to work, rather the algorithm decides whether the input was a pulse or white noise from which we may decide if the excitation was voiced or unvoiced. Real speech is more complicated than just voiced or unvoiced and the excitation may be both pulsatile and noise-like. Nevertheless, the algorithm is an improvement over previous algorithms in that it can estimate the input and the tract parameters.

The algorithm may be shown to be stable and to provide a unique solution following the method given in Appendix III of (Miyanaga et al., 1982). Several factors affect the convergence of the WRLS-VFF-VT algorithm: 1) model order, 2) stationarity of the signal, and 3) size of the data analysis interval. We have assumed that the model order may be determined a priori and that the data analysis interval is sufficiently large for the algorithm to work. From Eqs. (3 − 41) and (3 − 42) one can show that

$$e_k{}^2 = r_k{}^2 \ (1 + \lambda_{k-1}^{-1} \ [\hat{\phi}_k{}^t \ P_{k-1} \ \hat{\phi}_k])^2 \qquad (3-51)$$

If 1) the covariance matrix P_{k-1} is positive definite and 2) $[\phi_k{}^t P_{k-1} \phi_k]$ converges to zero as k goes to infinity, then the variance of the prediction error, ε_k, converges to the variance of the residual error, r_k. These two conditions can be shown to be satisfied for a stationary ARMA process with white noise, zero mean excitation (Makoul, 1976; Miyanaga et al., 1982; Ting, 1988). For this situation λ_k is near unity. When the excitation is not white noise, but a pulse train, then λ_k should be small to allow the detection of the excitation. In addition, a closed phase analysis may be used to ensure that the data can be modeled as an AR or ARMA process.

Table 3−1.

Adaptive WRLS-VFF-VT Algorithm with Input Estimation

Prediction error: $\xi_k = y_k - \hat{\phi}_k^{\,t}\,\hat{\theta}_{k-1}$

Gain: $K_k = P_{k-1}\,\hat{\phi}_k\,[\lambda_{k-1} + \hat{\phi}_k^{\,t}\,P_{k-1}\,\hat{\phi}_k]^{-1}$

Forgetting Factor: $\lambda_k = 1 - \xi_k^{\,2}\,(1 - \hat{\phi}_k^{\,t}\,K_k)^2 \,/\, E_1$

Threshold: $E_k = 1/M \sum_{i=0}^{M-1} \lambda_{k-i}$

If $E_k < 0.9$, then $\lambda_{thr} = 0.99 * E_k$

If $E_k > 0.9$, then $\lambda_{thr} = 0.9 * E_k$

If $\lambda_{thr} > \lambda_{min}$, then $\lambda_{thr} = \lambda_{min}$

Input estimate:

a) If $\lambda_k < \lambda_{thr}$, then the input is a pulse and

$\hat{u}_k^{\,w} = 0$

$\hat{u}_k = \hat{u}_k^{\,p}$

$= y_k - \hat{\phi}_k^{\,t}\,\hat{0}_{k-1}$

b) If $\lambda_k > \lambda_{thr}$, then the input is white noise and

$\hat{u}_k^{\,p} = 0$

$\hat{u}_k = \hat{u}_k^{\,w}$

$= \xi_k\,(1 - \hat{\phi}_k^{\,t}\,K_k)$

Real prediction error:

$e_k = (y_k - \phi_k^{\,t}\,\hat{\theta}_{k-1} - \hat{u}_k^{\,p})$

Parameter: $\hat{0}_k = \hat{\theta}_{k-1} + K_k\,e_k$

Covariance matrix: $P_k = \lambda_k^{\,-1}\,[P_{k-1} - K_k\,\hat{\phi}_k^{\,t}\,P_{k-1}]$

The performance of the WRLS-VFF-VT and the algorithms mentioned above were evaluated based on 1) the total formant/antiformant estimation error and 2) the rms spectral estimation error. The following items were used to calculate the estimation error.

(1) pole-zero estimation

(2) formant/antiformant tracking

The estimation error of the pole and zero frequencies (formant/antiformant) are defined as

$$E_p = \sum_{i=1}^{p} (\Delta F_{pi}/F_{pi}) \qquad (3-52)$$

and

$$E_q = \sum_{i=1}^{q} (\Delta F_{zi}/F_{zi}) \qquad (3-53)$$

where p and q are the number of poles and zeros, respectively, F_{pi} is the frequency of the ith pole, ΔF_{pi} is the estimation error of that frequency, and similarly for F_{zi}. The pole and zero frequencies were determined by polynomial root factoring.

The magnitude-square spectra of the AR and ARMA models, respectively, can be obtained by

$$|H(w)|^2 = G^2/|1 + \sum_{k=1}^{p} a_k \exp(-jw_k)|^2 \qquad (3-54)$$

$$|H(w)|^2 = \frac{b_0^2|1 + \sum_{k=1}^{q} b_k \exp(-jw_k)|^2}{|1 + \sum_{k=1}^{p} a_k \exp(-jw_k)|^2} \qquad (3-55)$$

The rms spectral error can be obtained using (3 – 54) and (3 – 55), namely

$$E_{rms}\ (dB) = \left\{ \left(\frac{1}{N}\right) \sum_{i=0}^{N-1} \left[10\log|H(w_i)|^2 - 10\log|\hat{H}(w_i)|^2 \right]^2 \right\}^{1/2} \qquad (3-56)$$

3.2.4 Complexity

The WRLS-VFF algorithm requires on the order of $(5(p+q)^2 + 6(p+q))$ floating point multiplications and additions (flops) per data point (Ting, 1989). Block data processing techniques used for AR estimation are generally less complex, e.g., Durbin's algorithm requires only (p^2+pN) flops for N data samples and Burg's lattice algorithm requires $((5/2)p^2 + pN)$ flops. (Most adaptive least squares algorithms can reduce the computational complexity to O(pN).) In comparison the WRLS-VFF algorithm, as implemented in Table 3–1, requires $(5Np^2 + 6Np)$ flops for an AR process. However, as noted below, the algorithm can be implemented with O(Np) for an ARMA process.

For ARMA parameter estimation, conventional block data processing techniques such as the Yule-Walker equation approach (LSMYWE) require on the order of $((1/6)p^3 + Np^2/2 + Np)$ flops. The WRLS-VFF algorithm requires $(5N(p+q)^2 + 6N(p+q))$ flops. The chief contributor to computational complexity for the WRLS-VFF algorithm is the computation required for the gain vector, $P_k\hat{\phi}_k$, which requires on the order of $(p+q)^2$ flops per data point. By using the idea of shift low rank the WRLS-VFF algorithm can be implemented with $O(N(p+q))$ flops instead of $O(N(p+q)^2)$ (Lee et al., 1981). Consequently, the WRLS-VFF algorithm can be made the same order of complexity as other non-adaptive or other recursive algorithms.

CHAPTER 4
ANALYSIS I: V/U/M/N/S (FIVE-WAY) CLASSIFICATION AND PITCH PERIOD DETECTION

Classification of the speech signal into voiced, unvoiced, mixed voiced, nasal, and silent (V/U/M/N/S) regions is usually the first step in speech analysis. Systems which do this classification using the speech signal range from simple ones that threshold the short-time energy and measure the zero crossing rates to systems using sophisticated pattern recognition techniques (Rabiner, 1978; Markel, 1976).

Accurate pitch estimation is essential since errors in this measurement can have a significant effect on speech analysis at later stages. Deviations of F0 greater than 1 % can be perceived by experienced listeners (Laver, 1980).

Natural voiced speech has period-by-period interval perturbations and amplitude variations. These factors must be measured and used in speech reproduction if a high-quality voice replication is to be achieved.

The new pitch detection algorithms will be discussed in section 4.1. The three pattern classification methods for making the V/U/M/N/S decision and features for use in V/U/M/N/S classifiers are discussed and described in section 4.2. The pitch-synchronous V/U/M/N/S algorithm will be discussed in section 4.3. In the last section, experimental results are discussed.

4.1 Pitch Detection Algorithms

Pitch determination, viz. the detection and measurement of voice fundamental frequency (f_o) in natural human speech, still constitutes one of the most

integral and at the same time most problematic areas of speech analysis. It has been established in numerous studies and even more so in practical computer speech applications that the accurate representation of the voicing characteristics is of paramount importance for all aspects of speech signal processing (synthesis, coding, transmission, compression, enhancement, etc). In speech coding, for instance, the quality of the vocoded speech deteriorates rapidly as a function of imprecise pitch estimates. In speech synthesis, considerable effort is dedicated today to the development and implementation of prosodic models for the generation of natural sounding pitch contours in text-to-speech.

Numerous studies have been dedicated to the design and evaluation of literally hundreds of pitch determination algorithms. Most of these studies have focused on pitch determination in stationary, quasi-periodic speech signals. Nonstationary, aperiodic signals, i.e. exhibiting occasional time and/or amplitude irregularities between successive periods of glottal excitation, have traditionally been disregarded and considered as reflecting pathological voice phenomena which are of no immediate concern to normal speech processing applications. However, aperiodic glottal vibrations occur far too often in normal, nonpathological voices of both women and men, to be classified simply as a clinical syndrome or voice disorder. Aperiodic phonation has also been shown to be systematically employed by human speakers as an important demarcation cue in the continuous speech utterance. Standard pitch detection and pitch estimation algorithms usually fail to correctly identify patterns of aperiodic voice excitation, which are often wrongly classified either as voiceless stretches of speech, or associated with faulty pitch values, typically of the "octave error" type.

We can divide the speech based pitch determination algorithms (PDA) into two categories: time domain PDAs, such as the SIFT method, and short term PDAs, such as the autocorrelation and cepstrum methods. The time domain PDAs are

capable of providing accurate pitch periods, but are sensitive to signal degradations in the analysis frame. Short-term PDAs, on the other hand, are more robust, but provide only an average pitch estimate over a number of periods, because short-term PDAs rely on the similarity of the speech signal over adjacent pitch periods. Thus, if a large number of pitch periods are contained in an analysis segment the estimated pitch value is a "smeared" or average value for the segment. Both methods lose all information about the absolute position of the glottal excitation.

Three methods were studied for the pitch estimates on a period by period basis in this study. First, we studied an EGG based method. Krishnamurthy and Childers (1986) proposed the EGG based pitch detection technique. The EGG is a periodic signal with two zero crossings per period during the voiced segments. The pitch period can be estimated as the time duration between two successive "invariant" features in the EGG. For example, either the positive or negative zero crossings or the sharp negative spike in the differentiated EGG at glottal closure can be used for this purpose with a simple threshold. However, a simple threshold may give the missing of the peak position in the beginning of a sentence and in the mixed sounds and in the weak sounds. Moreover, for a pathological speech mode such as vocal fry which has two or three opening/closing pulses in a glottal cycle, it remains difficult to detect the pitch period with the EGG signal. To solve this difficulty, the variable threshold was used in this study.

Secondly, we studied the LP error signal based method. Markel (1972, 1973) proposed a pitch estimation method that employs the autocorrelation function of the LP error signal. This method is capable of providing accurate pitch periods, but loses all information about the absolute position of the glottal excitation. Childers et al. (1991) developed the two-pass method for accurate, automatic glottal inverse filtering. This method first identifies the location of the main pulses of the LP error signal using a peak picking method with a simple threshold. The interval between the main pulses

of the LP error signal derived during the first pass of the inverse filtering procedure gives the pitch period. However, the peak picking method with simple threshold may miss the location of the main pulses of the LP error signal. The variable threshold was used in this study.

In the last method, we show that the VFF based method using WRLS-VFF-VT algorithm can accurately estimate and track the pitch period and the glottal closed phase intervals. The VFF can be updated recursively at each sample.

These methods are the time domain PDAs, and provide the pitch on a period by period basis, and provide a robust pitch detector. Figure 4−1 shows a speech waveform, an LP error signal derived from a fixed-frame LP analysis, the synchronous differentiated EGG signal and the VFF signal. The peaks in the linear prediction (LP) error and VFF signals occur nearly simultaneously with the negative peaks of the DEGG signal, which correspond to the instants of glottal closure. From these observations, we developed the "EGG based method", "LP error based method", and "VFF based method" for pitch detection on a period by period basis, and for the information of the absolute position of the glottal excitation.

A standard time domain pitch determination algorithm, a SIFT algorithm, is used to evaluate for their capacity to accurately detect and identify these patterns of aperiodic/quasi-periodic voice vibration in natural human speech.

4.1.1 SIFT and Modified SIFT Algorithms for the Pitch Detection

Figure 4−2 shows a block diagram of the SIFT method of pitch detection. A block of 400 samples (40 ms at a 10kHz rate) is low-pass filtered to a bandwidth of 900 Hz, and then decimated (down sampled) by a 5 to 1 ratio. The coefficients of a 4th-order inverse filter are obtained using the autocorrelation method of LPC analysis. The 2kHz speech signal is then inverse filtered to give a spectrally flattened signal

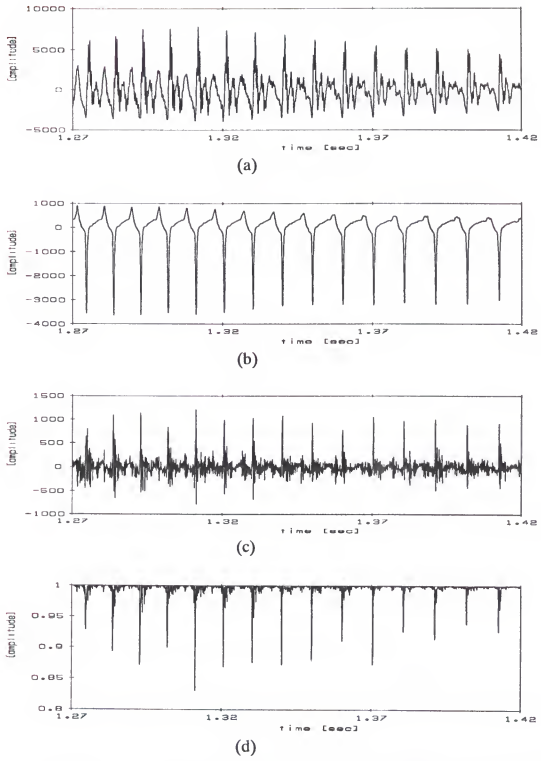


Figure 4–1. (a) Speech waveform, (b) synchronized DEGG, (c) LP residual error, and (d) VFF signal.

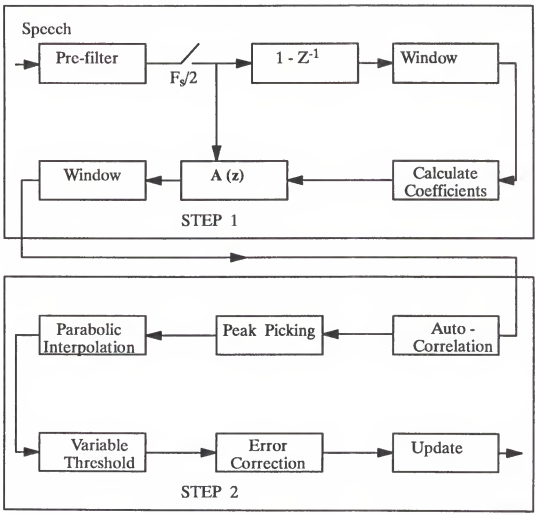


Figure 4−2. Block diagram of the SIFT algorithm (Markel, 1972).

which is then autocorrelated. The pitch period is obtained by interpolating the autocorrelation function in the neighborhood of the peak of the autocorrelation function (Markel, 1972). A V/U/M/S algorithm explained in section 4.2 is used for the error detection and correction of the starting and ending points.

Figure 4−7 (a) show the results of pitch detection using the SIFT algorithm. Two major errors occur. The first type of pitch error is the pitch period doubling or tripling or halving. The second type of pitch error is in the ending of the sentence where a kind of aperiodicity occurs between successive periods of glottal excitation. Nonstationary, aperiodic signals, occur far too often in normal, nonpathological voices of both women and men.

To correct the pitch period doubling or tripling or halving, a variable analysis frame length for peak picking is defined in this study. As the pitch period doubling or tripling is occurring, the analysis frame length for peak picking becomes smaller. As the pitch period halving is occurring, the analysis length increases. Two frames of delayed pitch information are retained for the pitch doubling or halving error detection and correction.

Figure 4−7 (b) show the results of pitch detection using the modified SIFT algorithm. Using the variable analysis frame size, it is able to correct pitch period doubling or tripling or halving errors in pitch detection and considerably improve the performance of a pitch detector. However, the errors in areas of the aperiodic glottal vibration occurring in the end of sentence cannot be corrected.

4.1.2 EGG based Algorithm

The electroglottographic (EGG) signal makes it easier to locate the glottal closed phase than is possible with the speech signal only (Krishnamurthy and Childers, 1986; Childers et al., 1990). A synchronized, differential EGG (DEGG) signal is used

to locate the closed glottal phase and to detect the pitch period. Figure 4–3 shows an example of EGG and DEGG waveforms with the synchronized speech signal. As can be seen in this figure, it is not difficult to obtain the pitch period, the glottal opening and closing instants from the EGG signal. While the glottal opening occurs relatively slowly, glottal closure is associated with a rapid reduction in tissue impedance and thus shows a large negative excursion in the EGG signal.

The estimation of the pitch period is simple using the EGG. The EGG is a periodic signal with exactly two zero crossings per period. The EGG period is also directly a result of the periodicity of vocal fold vibration. Thus, the pitch period can be estimated as the time duration between two successive invariant features in the EGG. For example, either the positive or negative zero crossings can be used for this purpose. Another choice is the sharp negative spike corresponding to glottal closure. One advantage of the EGG based pitch detection scheme is obvious; namely, the method computes the pitch period on a period by period basis. Consequently, the pitch smearing effects inherent in speech based methods are entirely avoided. The locations of the glottal closing instants determined from the EGG allow the isolation of individual periods of the speech waveform from closure to closure. Within each such period, the location of the opening instant is used to further divide the speech signal into closed and open glottis segments. Such fine segmentation of the speech waveform is essential for some of the speech analysis algorithms discussed later.

Figure 4–4 shows the block diagram of the EGG based pitch detection algorithm. The EGG varies in amplitude across speakers and across an utterance by an individual speaker. Consequently, a fixed threshold of the amplitude is generally inadequate to detect the location of the glottal opening/closing points.

The EGG signal is divided into frames, each frame consisting of 300 points. Successive frames overlap by 200 points. The maximum EGG amplitude and the EGG zero crossing rate in the frame are used to classify the frame as voiced or unvoiced. In

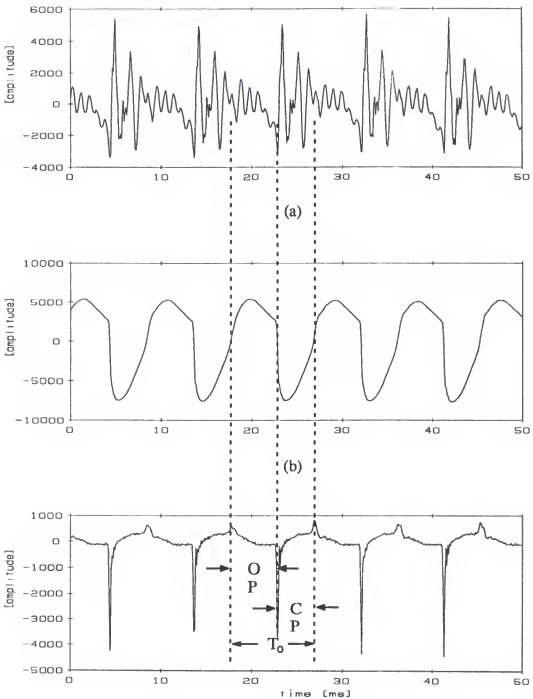


Figure 4−3 (a) Speech signal, (b) EGG signal, and (c) differentiated EGG (DEGG) signal.
* T_o: pitch period, OP: open phase, CP: closed phase

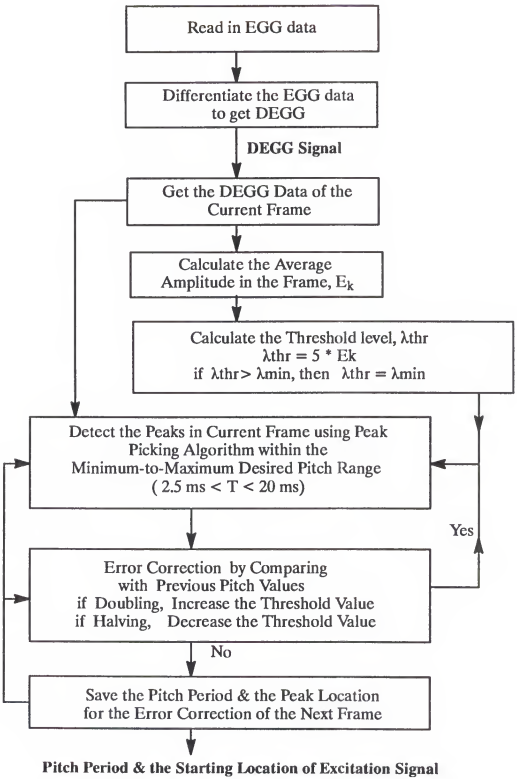


Figure 4−4. Block diagram of EGG based pitch detection algorithm.

the voiced frames, the average amplitude in a current frame is computed for the variable threshold of the peak picking using the DEGG. The opening and closing instants in the frame are located using the DEGG. Two frames of delayed pitch information are retained for the pitch error detection and correction.

The EGG signal is immune to surrounding acoustic disturbance, and provides a robust pitch detector. Moreover, the EGG, as we mentioned earlier, can also be used to isolate individual periods of the speech waveform such as the interval of glottal closure.

4.1.3 LP error based Algorithm

It is known that the LP error signal for a voiced speech waveform is characterized by peaked pulses separated by the pitch periods (Atal and Hanauer, 1971; Markel, 1972, 1973). These peaked pulses represent the main excitations to the LP vocal tract filter. As can be seen in Figure 4−1, the main excitation pulses in the LP error function consistently match the negative peaks of the DEGG signal, which were located very close to the instants of glottal closure (Childers et al., 1983; Krishnamurthy, 1983).

A block diagram of the LP error based method is shown in Figure 4−5. A pitch-asynchronous (fixed frame) LP analysis is performed on the input speech signal. For a voiced speech signal, the LP error function is characterized by a pulse train with the appropriate pitch period. The locations of these pulses are detected by a peak-picking method with variable threshold and are used as indicators of glottal closure.

The LP residual error signal is divided into frames, each frame consisting of 100 points. The analysis frame is overlapped by 100 points of previous and next frames, thereby totally 300 points. In the voiced frames, the average amplitude in a current

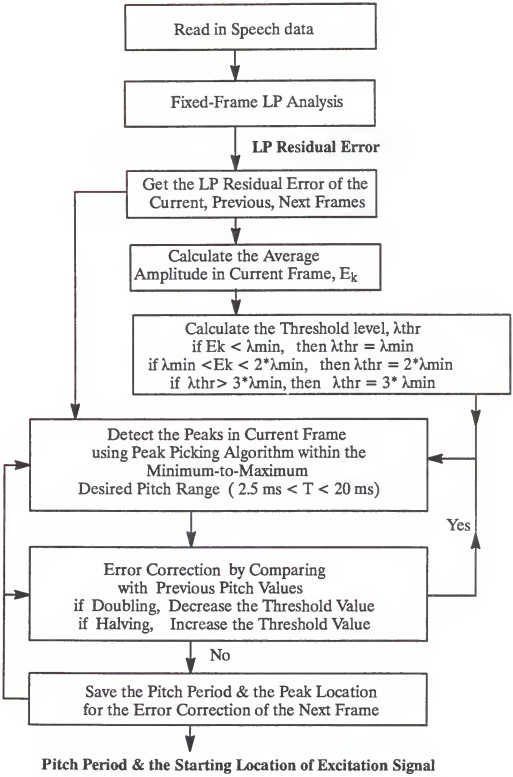


Figure 4−5. Block diagram of LP error based pitch detection algorithm.

frame is computed for the threshold of peak picking. Two frames of delayed pitch information are retained for the pitch error detection and correction. A V/U/M/S algorithm is also used for the error detection and correction of the starting and ending of a phrase or sentence. The LP error based method for pitch detection is capable of providing the pitch on a period by period basis with no extra auxiliary signal (EGG).

4.1.4 VFF based Algorithm

The VFF based method uses the VFF to identify the glottal closure points, which correspond to the instants of occurrence of the main excitation pulses for voiced speech. The smallest value of λ_k in the WRLS-VFF-VT method consistently matches the negative peaks of the DEGG signal, thereby indicating the instant of glottal closure in Figure 4−1.

A block diagram of the VFF based method for the pitch information is shown in Figure 4−6. During the WRLS-VFF-VT analysis, the λ_k can be obtained sequentially (sample-by-sample). The location of the excitation pulses and the estimation of the pitch period are detected by using a threshold value of λ_{min} and comparing it with each λ_k.

A variable threshold is used in the same way as in the LP error based method. If a peak crosses the variable threshold, its location becomes the pitch period candidate. Otherwise the frame is defined as unvoiced (i.e., pitch period = 0). An attempt at error correction is made by changing the threshold of the pitch period. The pitch period estimation occurs for the range 2.5 msec to 15.5 msec. In the final stage of pitch detection, V/U/M/S classification algorithm was used for the accurate pitch in the area of the beginning and the ending of the voice parts of a phrase or sentence. The VFF based method for pitch detection is capable of providing the pitch on a period by period basis with no extra auxiliary signal (EGG).

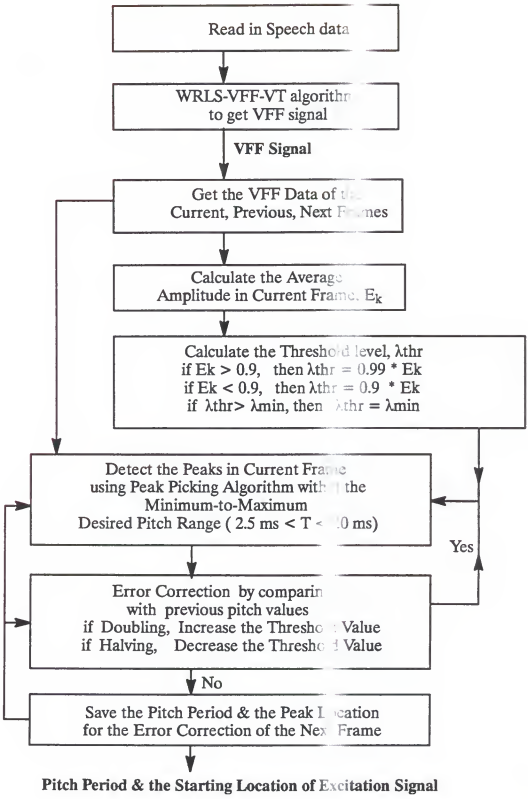


Figure 4−6. Block diagram of VFF based pitch detection algorithm.

4.1.5 Performance Evaluation

In order to evaluate the robustness of our pitch detectors, we compare the results for pitch estimation for the speech based SIFT algorithm with our methods in natural human speech as well as in pathologic human speech. We may use the EGG based method as a reference.

Figure 4−7 shows the pitch contours for the utterance "We were away a year ago" spoken by a male speaker. Figure 4−7 (c)-(e) show the results of the proposed methods which are the EGG based, LP error based, and VFF based algorithms, respectively. The pitch contours estimated from the proposed three methods are virtually identical and show the fine detail of the intonation changes. However, the result for the SIFT algorithm in Figure 4−7 (a) has the error caused by the pitch period doubling or tripling or halving, and also missed some fine detail of the intonation changes. Using the modified SIFT algorithm, the errors of the pitch period doubling and halving were corrected as shown in Figure 4−7 (b). But the modified SIFT algorithm did not perform well for the aperiodic voice vibration contained in the end of the sentence.

A pathological speech mode, vocal fry, is exemplified in Figure 4−8. For the vocal fry the vocal fold length is short. Vocal fold thickness varies with F_0, being medium for modal voice, while the vocal folds become thick for vocal fry and thin for falsetto (Hllien et al., 1968; Hollien and Colton, 1969; Boone, 1971; Allen and Hollien, 1973). For the vocal fold vibratory pattern, vocal fry is characterized by a glottal area function that has sharp, short pulses followed by a long closed glottal interval. The glottal opening phase may have one, two, or three opening/closing pulses (Moore and von Leden, 1958; Timcke et al., 1959; Hollien, 1974). The vocal fry EGG waveform also shows the double opening/closing pattern during an individual glottal cycle, as observed by other researchers(Moore and von Leden, 1958; Timcke et al., 1959;

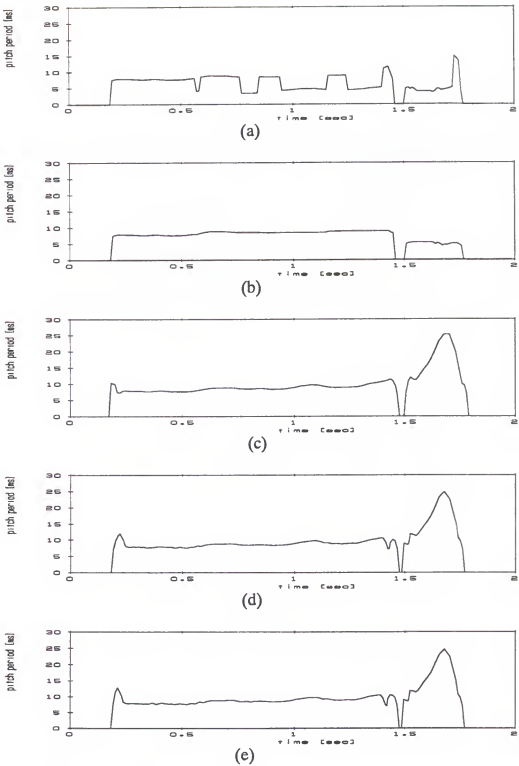


Figure 4−7. Pitch contours for a sentence, "We were away a year ago." using: (a) SIFT, (b) modified SIFT, (c) EGG based, (d) LP error based, and (e) VFF based methods.

Whitehead et al., 1984; Klatt and Klatt, 1990). Krishnamurthy and Childers(1986) reported that vocal fry is a speech mode that remains difficult to detect even with the aid of the EGG. This is due to multiple zero crossings of the EGG signal within one pitch period. Eskenazi (Eskenazi, 1988; Eskenazi et al., 1990) failed to obtain the pitch period using EGG signal for the pathologic speakers with vocal fry.

The speech, EGG, and DEGG for the sustained vowel,'i', spoken by a vocal fry speaker, which was used by Eskenaze (1988), are shown in Figure 4−8. We can find two or more glottal opening phases within one pitch period from EGG and DEGG signals.

The pitch contours from four algorithms are shown in Figure 4−9. In Figure 4−9 (a), the contour derived from the modified SIFT algorithm is smooth, but some fine detail is missed. Unlike the proposed methods, the modified SIFT algorithm uses an average pitch estimate over several periods, so the result provides the similarity over the adjacent pitch periods. The result from EGG based algorithm in Figure 4−9 (b) shows the details of the intonation and two abrupt changes in the start and in the end of the signal. This is caused by the incomplete glottal opening/closing in the beginning and ending areas. The LP error based and the VFF based algorithms produce contours that have ripples, spurious values in the middle of contours, and the fine detail of the intonation changes as shown in Figure 4−9 (c) and (d).

We have also analyzed an extensive normal data set used in the V/U/M/N/S classification algorithm discussed in section 4.2. as well as a pathological data set. It can be seen from this data that all of the proposed pitch detection algorithms perform well with respect to irregularity in pathologic speech signals.

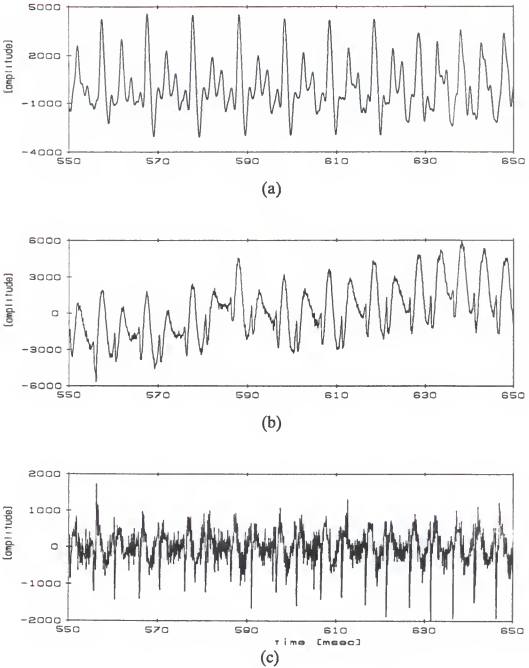


Figure 4–8. Pathological signals, vocal fry: (a) speech, (b) EGG, and (c) DEGG.

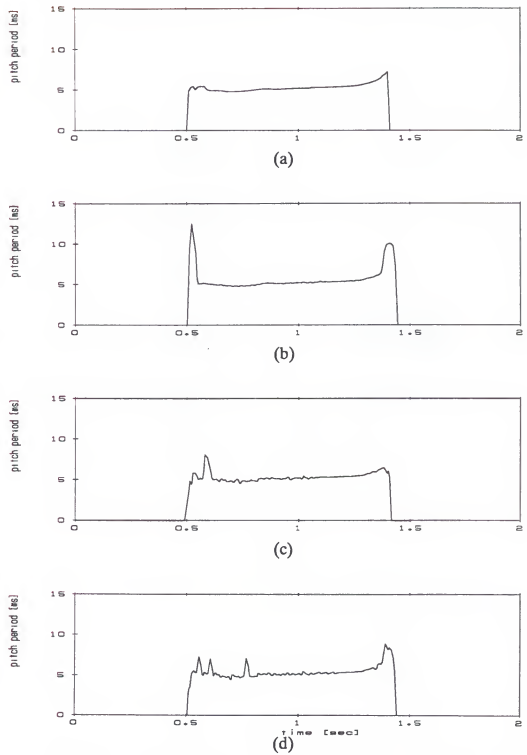


Figure 4–9. Pitch contours from four algorithms for sustained vowel "i" spoken by a pathologic speaker, vocal fry: (a) modified SIFT, (b) EGG based, (c) LP error based, and (d) VFF based.

4.1.6 Summary

If the EGG based method is considered as a reference, we can therefore conclude that our pitch detectors are very reliable in quasi-periodic as well as in aperiodic speech signals. The "pitch smearing" effect inherent in speech signal based methods is avoided, and pitch values are available on a period by period basis. This can have important applications for several problems where accurate pitch contour estimates are desired (Hess, 1982; Lebrum, 1971).

4.2 One-Channel Five-Way Classification

In previous work, a two-channel two-way (V/U-S) (Krishnamurthy, 1986) and a two-channel four-way (V/U/M/S) (Childers et al., 1989) classification algorithms were described for automatically classifying speech. These algorithms used the speech and electroglottogram (EGG) signals. We recognize that in many situations, the EGG signal is either unavailable or cannot be used.

In this chapter, we describe the one-channel-based algorithm, for the V/U/M/N/S classification. The decision making process is viewed as a pattern recognition problem. Two aspects of the task are tested: feature selection and classifier type. The feature selection procedure is done for identifying a set of features to make V/U/M/N/S classification. The classifiers used are a rule-based tree-structure method (so called "decision tree method"), a vector quantization (VQ), and a neural network (NN).

4.2.1 Introduction

In the most commonly used model of speech production, the speech signal is decomposed into a filter component and an excitation component. The excitation component is represented by one of two states: voiced-more or less periodic, produced by vibration of the vocal cords, or unvoiced-noise like, produced by forcing air past some constriction in the vocal tract. Using this model, considerable success has been achieved by employing pattern classification techniques to assign a segment of speech to one of two classes, voiced or unvoiced (Atal and Rabiner, 1976; Siegel, 1979). Despite the widespread use of this simplified model, the restriction of the excitation to the two classes is not adequate for the synthesis of high quality speech from analysis parameters. Experiments show that high quality synthesis requires mixed excitation for synthesis of the voiced fricatives (v(vote), th(then), z(zoo), z(azure)). Human production of these sounds involves the vibration of the vocal cords in conjunction with a turbulent air flow at some point of constriction. Speech synthesizers driven from stored data rather than from analysis parameters commonly include a link between the unvoiced and voiced excitation paths to allow a mixed excitation in the synthesized speech. In order to allow a mixed source in an analysis-synthesis system, the excitation for a segment of speech must be identified as voiced, unvoiced, or a combination of voiced and unvoiced.

The acoustic structure of nasal consonants has long been predicted by the acoustic theory of speech production (Fant, 1960; Fujimura, 1962; Flanagan, 1972). The presence of a side-branching resonator (the blocked oral cavity) will introduce an antiresonance (zero) in the spectrum of nasal consonants. Theoretically, the antiresonance can be used to identify the place of articulation of nasal consonants because the frequency of the antiresonance is determined by the dimension of the side-branching resonator. Such differences, however, are difficult to detect using

conventional techniques of spectral analysis. The presence of the spectral zero introduces nonlinear equations to parametric methods of spectral analysis (Kay, 1987). Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator and anti-resonator into the cascade vocal tract model in the formant synthesizer. The Klatt synthesizer includes an additional resonance-antiresonance pair for synthesizing nasals. It is necessary to identify nasalized segments in the recorded speech in order to decide when this branch should be activated. Another purpose of nasality detection is the correction of the all-pole estimates of the formant frequencies and bandwidths; it is known that the presence of zeros in the spectra of nasal speech tends to move the formants upwards in frequency and to increase their bandwidths.

Therefore, to get the synthesis of high quality speech from analysis parameters, extending the V/U/S or the V/U/M/S decision to V/U/M/N/S decision is needed.

4.2.2 Features

The features considered for use in making the V/U/M/N/S classification can be divided into two categories: features for V/U/M/S decision and features for nasal/nonnasal decision.

4.2.2.1 V/U/M/S Decision

In previous work (Krishnamurthy, 1986; Childers et al., 1989), a two-channel, four-way classification algorithm was described. Instead of using both the speech and the EGG signal, the one-channel four-way classification algorithm utilizes only the speech signal. The EGG signal is usually unavailable in real situations and a designer

has to design a speech system that relies only on speech input. In this case, it is not possible to take advantage of the EGG signal as an indicator of vocal fold vibration and, as a result, the system becomes more complicated.

Time-domain analysis techniques, such as zero crossing rate, energy, and level crossing rate, are not sufficient to achieve a successful one-channel four-way classification. This is clearly shown by the ranges of such features, as shown in Table 4–1. Different features, such as spectral distribution (Rabiner et al., 1977) or the LP error signal (Rabiner and Schafer, 1978; Atal and Rabiner, 1976), have to be included in the feature set if a reliable classifier is needed. Hence, in this study, six spectral energy ratios and normalized autocorrelation coefficients are added to the time-domain features to form the feature set.

The same set of six sentences is used as the training data set to develop a one-channel four-way classifier as was used for our two-channel classifier (Childers et al., 1989). The following parameters (measurements) are computed for each block of samples:

1) Normalized log energy $SENG_s$ - defined as

$$SENG_s = 10 * \log_{10}\left[10^{-5} + \frac{1}{N}\sum_{n=1}^{N} s^2(n)\right] \qquad (4-1)$$

N: no. of samples in pitch interval

2) Normalized autocorrelation coefficient at unit sample delay, C_1, which is defined as

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\left[\sum_{n=1}^{N} s_2(n)\right]\left[\sum_{n=1}^{N-1} s_2(n)\right]}} \qquad (4-2)$$

3) Level crossing rate of speech signal, SLCR

4) Zero crossing rate of the differentiated speech signal, SDZCR

5) Zero crossing rate of the speech signal, SZCR

6) Ratio1: ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 3800 - 4200 Hz

7) Ratio2: ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 4200 - 4600 Hz

8) Ratio3: ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 4600 - 5000 Hz

9) Ratio4: ratio of the spectral energy of the speech frame in the 800 - 1200 Hz band to that in the 4200 - 4600 Hz

10) Ratio5: ratio of the spectral energy of the speech frame in the 800 - 1200 Hz band to that in the 4600 - 5000 Hz

11) Ratio6: ratio of the spectral energy of the speech frame in the 430 - 470 Hz band to that in the 4400 - 4800 Hz

As with the two-channel four-way classification (Childers et al., 1989), all the feature values are evaluated on a frame by frame basis with a frame size of 100 data points (10 milliseconds).

The spectrum of voiced sounds shows that most of the energy is concentrated below 1 kHz and the first formant, usually the highest peak, is located below 350 Hz. For unvoiced sounds, most of the speech energy is found above 2.5 kHz and the highest peak is also found in this region. (Even though the first formant for unvoiced sound is usually located below 450 Hz, its energy level is lower than that of the third or the fourth formant.) In the case of mixed sounds, the spectrum is relatively flat for the whole frequency region. The examination of the spectra of mixed sounds indicates that there are usually two peaks. One is located below 1 kHz and the other above 3 kHz. It is believed that the former is produced by the low frequency carrier component (due to

a vocal fold vibration) and the latter is caused by the noise-like high frequency component (due to a turbulent airflow), both of which exist in a mixed sound. In Figure 4–10, examples of spectra for voiced, unvoiced, mixed, and silent segments are shown.

4.2.2.1.1 Methodology for Spectral-Domain Features

Six spectral energy ratios were selected as features representing the spectral properties of the speech signal. The selection of six, rather than a single ratio as some other researchers have done, is based on the observations that 1) in the sentences pronounced by one speaker, a significant spectral deviation can occur even for one phoneme due to the phonetic environment, 2) the same phoneme spoken by one speaker at different times can have different spectral distributions according to the speaker's condition, mood, and intention (intra-speaker variability), and 3) for different speakers, the same phoneme can have considerably different spectral distributions (inter-speaker variability).

It is well known that the quality of the periodogram gets worse with increasing data length. Namely, the variance of the periodogram is proportional to the data length and a frame having more than 100 points usually results in a useless periodogram for speech scientists. Hence, in order to obtain a more consistent spectrum estimate, Bartlett (Openheim and Schafer, 1975) suggested a modified version of the periodogram evaluation technique. In his method, a frame to be analyzed was divided into smaller subframes, and the spectral estimate for each segment was evaluated as the convolution of the true spectrum with the Fourier transform of the triangular window function. The final spectral estimation was obtained by averaging the periodograms for the subframes. The variance of Bartlett's estimate decreased by the factor of the number of segments, and resulted in a consistent spectral estimate.

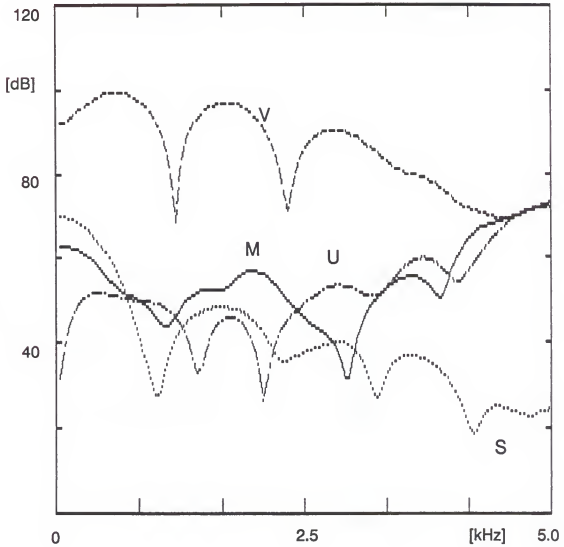


Figure 4−10. Spectral distribution of voiced, unvoiced, mixed, and silence.

Welch (Welch, 1967) has introduced a modification of the Bartlett procedure that is particularly well suited to direct computation of a power spectrum estimate using the FFT (Fast Fourier Transform). A data frame of length N is further divided into K segments having M samples. The window, $w(n)$, is applied directly to the data segments before computation of the periodogram. The modified periodogram of the i-th segment can then be defined as

$$J_{i,M}(\omega) = \frac{1}{MU}\left[\sum_{n=1}^{M-1} x_i(n)w(n)e^{-j\omega n}\right]^2 \qquad (4-3)$$

where

$$U = \frac{1}{M}\sum_{n=1}^{M-1} w^2(n) \qquad (4-4)$$

and the spectrum estimate is defined as

$$B_{xx}(\omega) = \frac{1}{K}\sum_{i=1}^{K} J_{i,M}(\omega) \qquad (4-5)$$

In this method, the variance of the final periodogram is also reduced by a factor of K.

In this study, the spectral distribution of the speech signal is evaluated with this Welch method and a Hamming window. The window size is 28 samples and five windows are fitted into a frame, which results in a 28.6% overlap between two adjacent windows. In order to improve the resolution of the periodogram, 228 zeros are appended before executing FFT to each windowed data set. The final resolution of our periodogram is 39.1 Hz.

4.2.2.1.2 Statistical Properties of Features

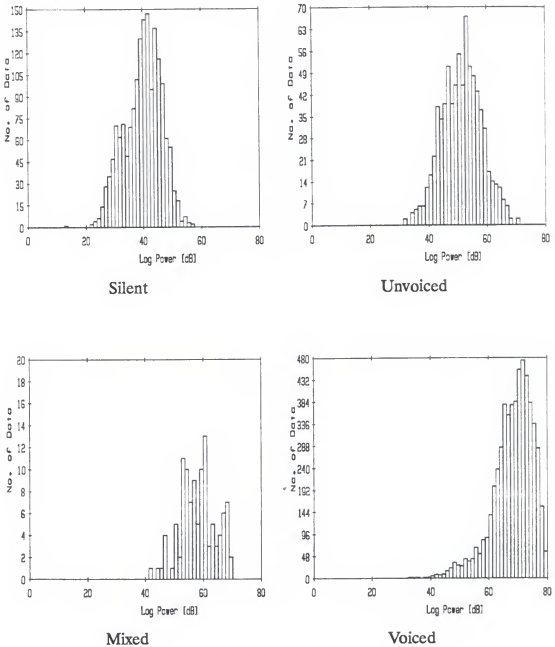
Before proceeding to a detailed discussion of the decision algorithm, it is worthwhile discussing the expected nature of variation of each of the above feature parameters for the four classes. Measurement distributions of these parameters for the different classes are shown in Figure 4−11.

The means and standard deviations for the four classes of a typical set created by manually segmenting natural speech into regions of silence, unvoiced speech, voiced speech, and mixed speech are shown in Table 4-1. Six speakers (three male and three female) were used in the set with each speaking five sentences.

These parameters are correlated with each other. These correlations vary between the parameters and between the classes. The decision algorithm discussed in the section 4.2.3 makes use of these correlations to optimally combine their contributions in differentiating between the classes.

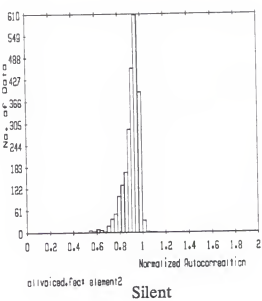
4.2.2.2 Nasal/nonnasal Decision

The search for invariant acoustic cues that indicate the presence of nasal murmurs in continuous speech has a long history. Fujimura (1962) reported the spectral characteristics of nasal murmurs in intervocalic contexts. He found three essential features: first, the existence of a very low first formant in the neighborhood of 300 Hz; second, the relatively high damping factors of the formants; and third, the high density of the formants in the frequency domain. Fant (1962) reported that a voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which the second formant is weak or absent; a formant at approximately 250 Hz dominates the spectrum but several weaker high-frequency formants occur, and the bandwidths of nasal formants are generally larger than in vowellike sounds.

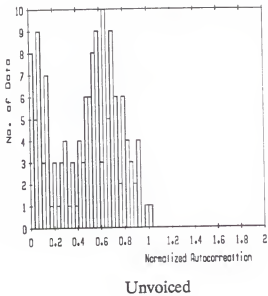


(a)

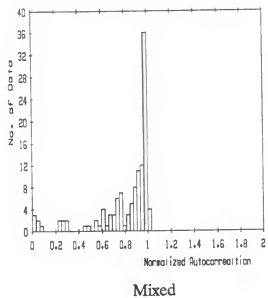
Figure 4−11. Measurement of density functions for (a) Log energy, (b) Autocorrelation coefficient, (c) Zero-crossing rate, (d) Level-crossing rate, (e) Differentiated zero-crossing rate, (f) Ratio1, (g) Ratio2, (h) Ratio3, (i) Ratio4, (j) Ratio 5, and (k) Ratio6.



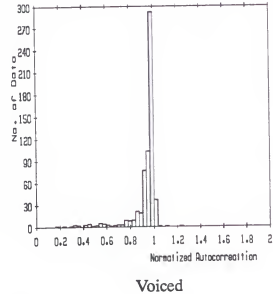
Silent



Unvoiced



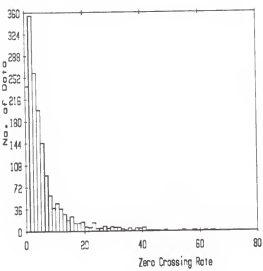
Mixed



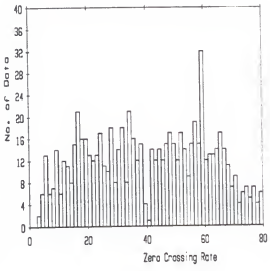
Voiced

(b)

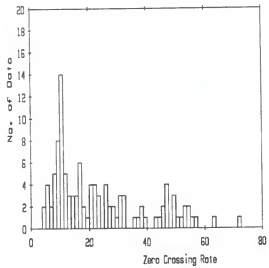
Figure 4−11. Continued



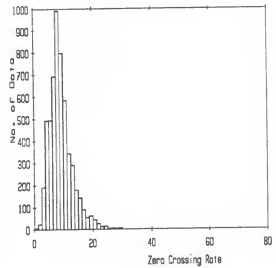
Silent



Unvoiced



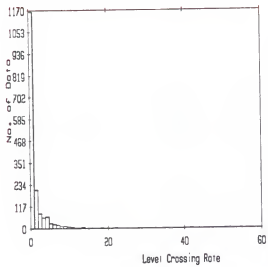
Mixed



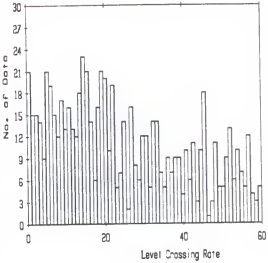
Voiced

(c)

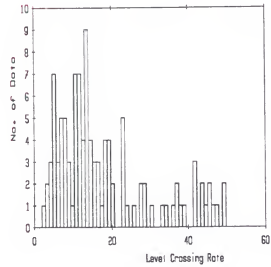
Figure 4−11. Continued



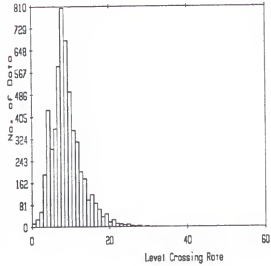
Silent



Unvoiced



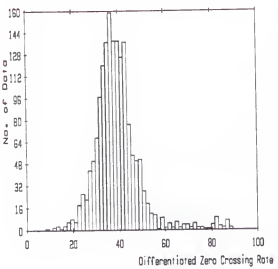
Mixed



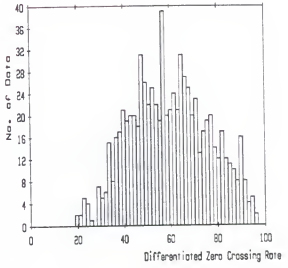
Voiced

(d)

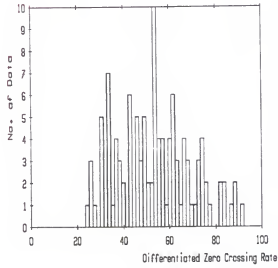
Figure 4−11. Continued



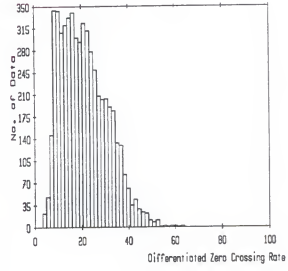
Silent



Unvoiced



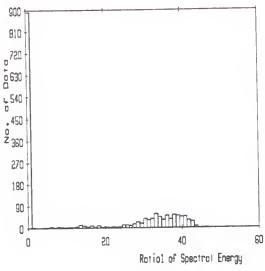
Mixed



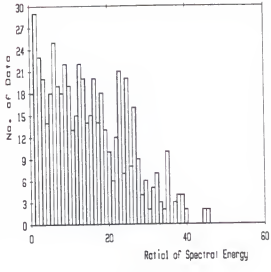
Voiced

(e)

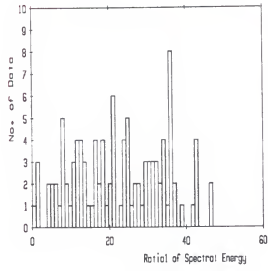
Figure 4−11. Continued



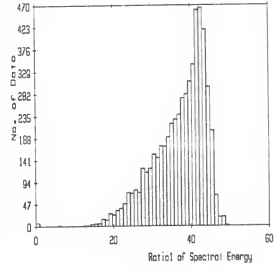
Silent



Unvoiced



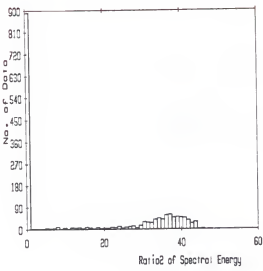
Mixed



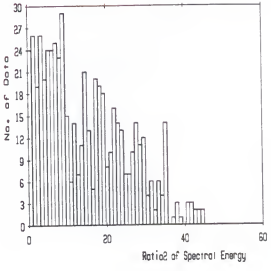
Voiced

(f)

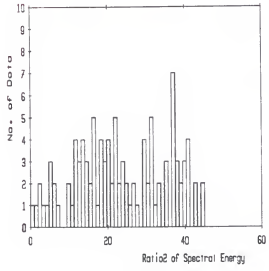
Figure 4−11. Continued



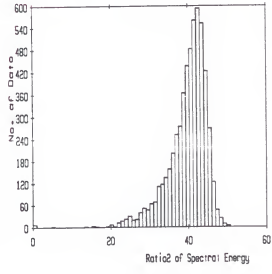
Silent



Unvoiced



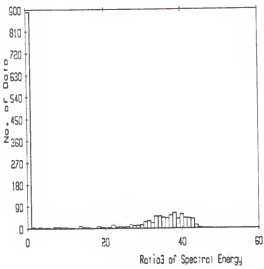
Mixed



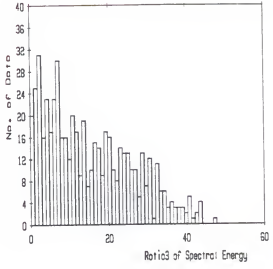
Voiced

(g)

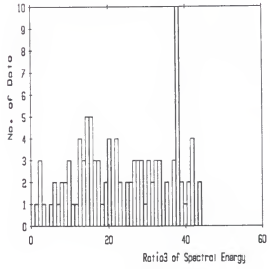
Figure 4−11. Continued



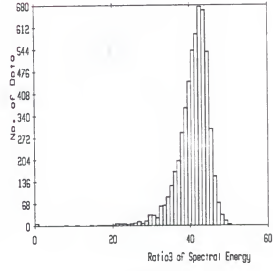
Silent



Unvoiced



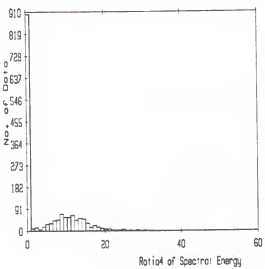
Mixed



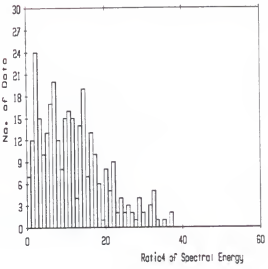
Voiced

(h)

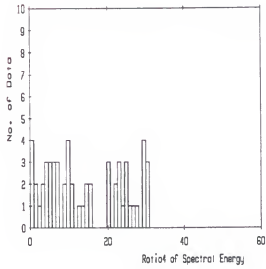
Figure 4−11. Continued



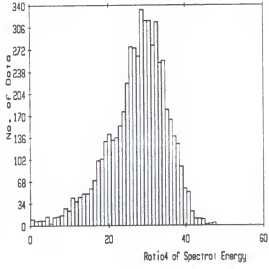
Silent



Unvoiced



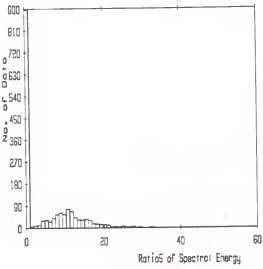
Mixed



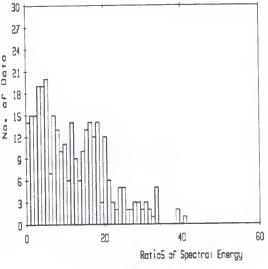
Voiced

(i)

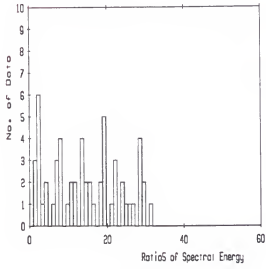
Figure 4−11. Continued



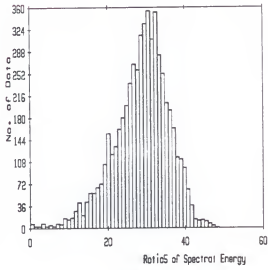
Silent



Unvoiced



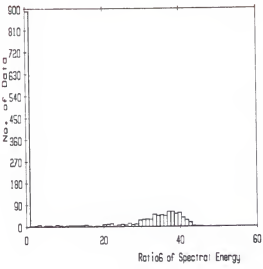
Mixed



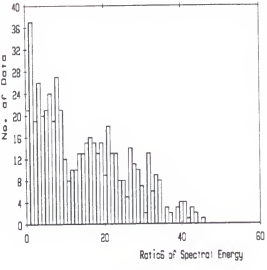
Voiced

(j)

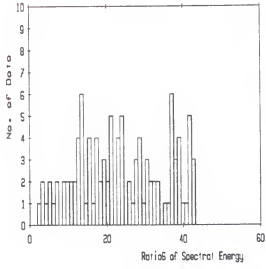
Figure 4−11. Continued



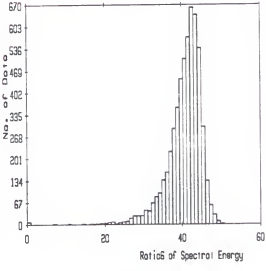
Silent



Unvoiced



Mixed



Voiced

(k)

Figure 4−11. Continued

Table 4−1 Means and standard deviation for the five classes for the training data described in section (6 - speaker, 5 - sentence)

	Log Power (dB)	Normalized Autocorr.	Level Crossing	Differential Zero-cross.	Zero Crossing
1) Silence					
Mean	36.55	0.9242	0.9036	36.89	6.072
Std. Dev	8.603	0.1503	1.795	9.063	7.754
1) Unvoiced					
Mean	51.2	0.007874	31.32	60.95	47.7
Std. Dev	6.853	0.5513	22.07	18.04	20.73
1) Mixed					
Mean	57.13	0.5671	16.46	51.26	24.78
Std. Dev	7.866	0.5325	16.33	23.82	21.34
1) Voiced					
Mean	66.24	0.9112	9.71	24.69	10.08
Std. Dev	7.657	0.07187	4.5	10.15	4.488

Ratio1	Ratio2	Ratio3	Ratio4	Ratio5	Ratio6
15.52	15.87	16.03	6.768	6.928	15.73
17.84	18.23	18.38	9.289	9.507	18.09
6.306	5.525	5.436	-4.286	-4.374	4.874
12.45	13.26	13.83	14.83	15.64	13.66
19.62	19.37	19.11	0.1937	-0.0724	18.75
16.19	17.2	17.86	13.64	14.32	17.58
38.37	42.14	44.11	24.23	26.2	43.45
5.877	3.585	2.103	5.515	5.426	2.325

Examination of spectrograms and spectral cross sections essentially confirmed Fujimura's report. A low-frequency nasal resonance and drop in mid-plus-high-frequency energy (above roughly 1000 Hz) in the absence of a significant drop in low-frequency energy (below 1000 Hz) were found to be reliable cues for nasals. However, when the same cues were tested on continuous speech, differentiation between nasals and nonnasals proved markedly poorer.

Several recent investigations of nasals(Kurowski and Blumestein, 1984, 1987; Repp 1986, 1988; Repp and Svastikula, 1988; Seitz et al., 1990), and of other consonants (Stevens and Blumstein, 1978; Kewley-Port, 1983; Lahiri et al., 1984; Forrest et al., 1988) have focused upon the interactions or integration of consonantal murmur or release and vowel transition signal portions in relation to human perception, as well as algorithmic classification of phonetic contrasts. In their works on nasals, they used a method for representing spectral change at the nasal-vowel boundary. Kurowski and Blumstein (1987) used a simple method for representing spectral change at the nasal-vowel boundary, in which they examined the proportion of the change in energy in the region of Bark 5-7 to the change in energy in the region of Bark 11-14 across two spectra, one computed for the two glottal pulses of the nasal murmur immediately preceding release, and one for the first two glottal pulses after release. Seitz et al. (1990) defined the spectral change from nasal to vowel in each of these bands as the maximum absolute difference obtained from the differencing; that is, as the difference in amplitude found at the frequency, within the 5-to-7 or 11-to-14 Bark band, where the vowel-minus-murmur difference is greatest. Spectra scaled in hertz were processed similarly, but the ranges used were 450-700 Hz (approximately equivalent to 5-7 Bark) and 1370-2150 Hz (approximately equivalent to 11-14 Bark). This procedure is slightly different from Kurowski and Blumsten's, which used the rms energy encompassing the low and high bands rather than the absolute difference maxima in these bands.

However, it is difficult to use these features in the classification of nasal/nonnasal on continuous speech directly. Our prime interest lay in classifying nasal/nonnasal in a variety of contexts such as may be encountered in free text rather than the perception of (m)-(n) distinction in CV syllables.

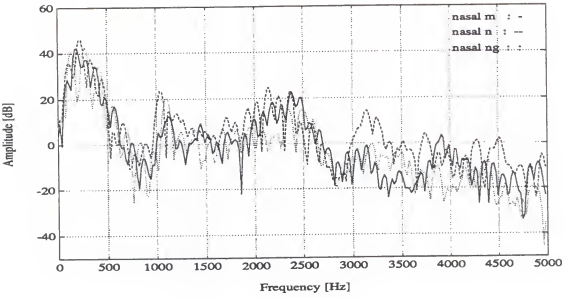
The nasal sound characteristics in the spectrum (Fujimura, 1962; Mermelstein, 1977; Nakata, 1959; Yea and Childers, 1983) may be summarized as follows:

1. The existence of a very low first frequency in the neighborhood of 300 Hz.
2. The bandwidth of nasal formants are generally larger than in vowel-like sounds.
3. The second formant is weak
4. The high density of the formants is in the frequency domain.

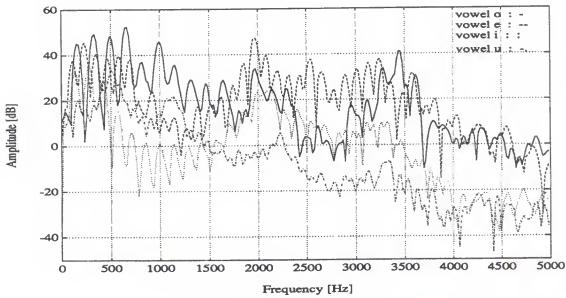
We have observed in some of our data that the first peak (around 250 Hz) in the spectrum of a nasal consonant is at least 40 dB greater than the first valley (around 800 Hz) in the spectrum. This is caused by the combination of the first and third characteristics of nasal consonants as described above.

An example of this feature is shown in Figure 4−12(a) for /m/, /n/, and /ng/ compared with various vowel sounds in Figure 4−12(b). This is the feature that may be of some practical value in the formulation of criteria for identifying nasal consonants as a class (Fant, 1960; Fujimura, 1962). According to Fujimura's paper (1962), this feature cannot be found in other vocalics (Fujimura, 1962).

The characteristics of time waveforms of nasal sounds is a low and smooth amplitude. It seems likely that an obstruction in the nasal airway may act to attenuate or absorb the sound waves as they pass through the nose resulting in a measurable decrease in the amplitude of the nasal phonemes and resulting in damping high



(a)



(b)

Figure 4−12. Characteristics in the spectrum for (a) nasals and (b) vowels

frequencies (Fant, 1960; Fujimura, 1962). This decrease is shown in the Figure 4−13 (a) for the real speech "nine" spoken by a male speaker.

As features representing the spectral properties, we employ four spectral energy parameters, all defined in relative values with respect to the energy in the first formant frequency band. Using the same bands tested by Seitz et al.(1990), spectral value in the 450-700 Hz (approximately equivalent to 5-7 Bark used by Kurowski and Blumstein (1987)) is subtracted from the values corresponding in the first formant frequency range, 150-400 Hz. The same was done for the 1370-2150 Hz (approximately equivalent to 11-14 Bark used by Kurowski and Blumstein (1987)). The ratios are evaluated from the spectral distributions obtained by applying the Welch method in the same way in 4.2.2.1.1. We define the ratios as following:

Ratio7: The ratio of the spectral energy in the 150-400 Hz band to that in the 450-700 Hz band

Ratio8: The ratio of the spectral energy in the 150-400 Hz band to that in the 1370-2150 Hz band

Ratio9: The ratio of the spectral energy in the 150-400 Hz band to that in the 2700-3100 Hz band

Ratio10: The ratio of the spectral energy in the 4600-5000 Hz band to that in the 1370-2150 Hz band

The zero-crossing rate was selected in the time domain feature for the reason that the zero-crossing rate of voiced sound has a larger value than that of nasal sound.

The results of applying this parameterization to the word,"nine", are shown in Figure 4−13 (d) - (g). One observes that all of the parameters show considerable separation by categories. The first two curves of this figure are the speech waveform and the spectrogram respectively. The Figure 4−13 (c) is the results of the V/U/M/N/S classification. This contour can assume one of five values where value 1 is silence, level 3 is unvoiced, level 5 is nasal, level 7 is mixed, and level 9 is voiced. It can be seen that the analysis made the nasal detection correctly.

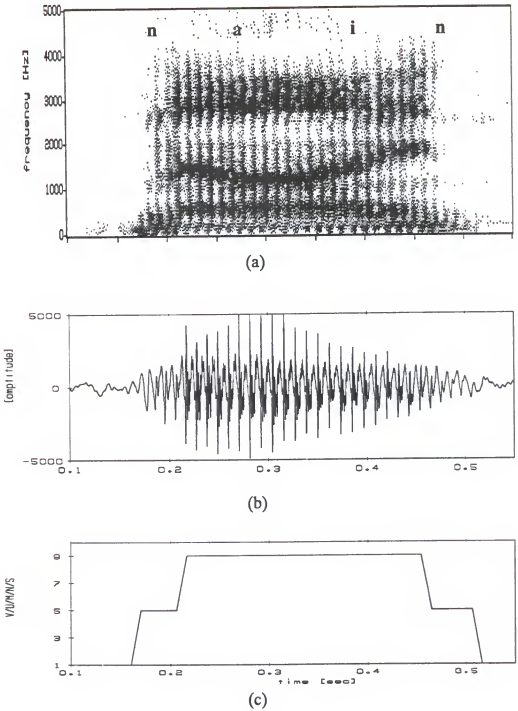
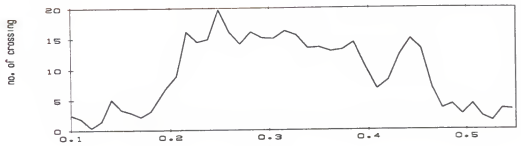
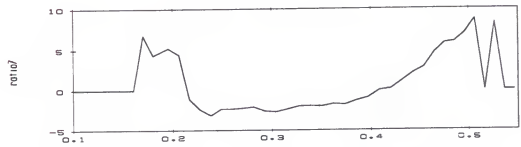


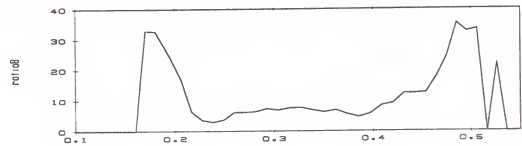
Figure 4−13. Characteristics of feature parameters in nasal/nonnasal detection for the real speech, "nine": (a) spectrogram, (b) waveform, (c) V/U/M/N/S result, (d) zero-crossing rate, (e) ratio7, (f) ratio8, and (g) ratio10.



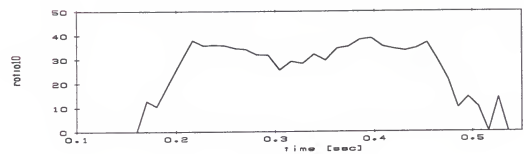
(d)



(e)



(f)



(g)

Figure 4−13 Continued

4.2.2.2.1 Statistical properties of features

The means and standard deviations for the two classes for a typical set created by manually segmenting natural speech into regions of nasal and nonnasal speech are shown in Table 4−2. Six speakers (three male and three female), each speaking five sentences, are used in the set. The nasals and nonnasal sounds are pooled and measured the means and standard deviation values for the measured parameters. Although there is a substantial overlap, the distribution of these parameters for the nasal/nonnasal classification have dominant peaks and have significantly different means.

The nasal spectrum depends on the color of the syllabic vowel because that is the underlying articulation on which the nasal murmur articulation is superimposed. Of course, the nasal spectrum further depends on the place of production. No attempts to use our measurements to categorize the nasal murmurs by pace of production have yet been carried out. Because good nasal/nonnasal classification is obtainable without consideration of place or production information, it appears appropriate for any complete analysis to do nasal/nonnasal classification first.

Our effort has focused on the effective combination of the information from several independently measured parameters and from the V/U/M/S classification results in an attempt to attain better classification performance.

Table 4–2. Means and standard deviations for the nasal sounds (m,n,ng) and the nonnasal.

	Zero Crossing	Ratio 7	Ratio 8	Ratio 9	Ratio 10
1) Voiced					
Mean	9.189	1.3	21.3	39.53	19.61
Std. Dev	3.986	2.447	6.9	4.674	7.091
2) Nasal 'm'					
Mean	5.444	3.954	28.99	40.02	15.38
Std. Dev	1.325	1.078	5.719	2.625	7.211
3) Nasal 'n'					
Mean	5.151	4.712	28.48	38.92	14.85
Std. Dev	2.078	1.417	7.202	3.726	7.246
4) Nasal 'ng'					
Mean	5.375	3.238	21.27	31.74	5.77
Std. Dev	2.621	2.117	14.24	10.32	6.02

4.2.3 Pattern Classification

No single feature seems to give consistently reliable performance in making the speech segmentation, so it is desirable to combine several features to obtain a good characterization of the segmentation of a speech signal. One way to incorporate a number of features is to view the segmentation decision process as a pattern classification problem. Atal and Rabiner (1976) have used a statistical model to design a minimum distance classifier for V/U/S classification. This requires assuming a particular distribution function for the features and computing the mean and covariance matrix for each class using a large enough set of data to obtain an accurate statistical characterization.

In this study, methods for making the five-way V/U/M/N/S classification decision are examined. Three classifier types are considered: vector quantization (VQ), neural network (NN), and decision tree method.

VQ is a process in which data to be encoded are broken into small "blocks" or vectors, which are then sequentially encoded vector by vector. The idea is to identify a set, or "codebook", of possible vectors which are representative of the information to be encoded. The VQ encoder pairs up each source vector with the closest matching vector from the code book, thus "quantizing" it. The decoding is a trivial matter of piecing together the vectors whose identity has been specified. The Pairwise Nearest Neighbor (PNN) algorithm is used, which is presented as an alternative to the Linde-Buzo-Gray (generalized Lloyd) algorithm to design a full-search VQ codebooks based on a training sequence of feature vectors is used (Linde et al., 1980).

For the NN method, the feed-forward multilayer back-propagation network was particularly well suited to the two-way V/U classification problem (Bendiksen, 1990). The principal advantages of this classification approach are its properties: 1) it focuses on correct classification of the difficult-to-classify "boundary" patterns, and 2) it makes no assumptions about the distributions of the features. The back propagation algorithm has been tested with a number of deterministic problems such as the exclusive OR problem, on problems related to speech synthesis and recognition and on problems related to visual pattern recognition. It has been found to perform well in most cases and to find good solution to the problems posed. Its principal disadvantages are computational.

The decision tree method uses a sequence of two-way decisions and has the potential advantage that the feature sets used for each discrimination can be selected independently for each decision. A second advantage is that the use of a sequence of binary classifications can allow a more flexible division of a feature space into five regions. Compared to pattern classification techniques which use one set of

discriminant functions to make the V/U/M/N/S decision, this approach allows greater flexibility in the decision surfaces which define the voiced, unvoiced, mixed, nasal, and silence speech regions.

For the V/U/M/N/S classifiers of VQ and NN methods two decision structures are combined: multiclass and binary. A decision-tree with the V/U/M/S decision at the root has been implemented. The subtrees of the root are voiced (V), unvoiced (U), mixed (M), and silence (S) segments. In the V subtree, the speech segments are classified as one of two terminal nodes, or leaves: nasal (N) and nonnasal voiced(V). The resulting tree has five leaves as shown in Figure 4−14: voiced (V), unvoiced (U), mixed (M), nasal (N), and silence (S). The advantage of this structure is that the feature set for V/U/M/S discrimination can be different from that for nasal/nonnasal discrimination.

4.2.3.1 Vector Quantization

4.2.3.1.1 Introduction

Vector quantization (VQ) was used for speech coding in the 1950s (Dudley, 1958) and was recently revived (Buzo et al., 1980; Linde et al., 1980; Markel et al., 1985). The technique has been used in speech recognition and speaker recognition (Furui, 1988; Pan et al., 1985; Soong et al., 1988; Soong et al., 1987). The VQ problem is part of the pattern recognition problem that is concerned with the classification of data into a discrete number of categories using a fidelity criterion. The basic concept of VQ, as applied to speech, is depicted in Figure 4−15, where $\mathbf{x_n} = (a_1, a_2, ..., a_N)^T$ represents an N-dimensional column vector whose components $\{a_k, 1 \le k \le N\}$ are real-valued random variables with continuous amplitude. The input vector $\mathbf{x_n}$ is mapped onto another real-valued discrete N-dimensional vector $\mathbf{y}$. Typically, $\mathbf{y}$ takes

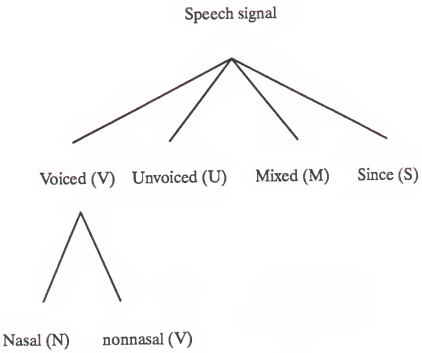


Figure 4−14. Decision structure in VQ and NN methods.

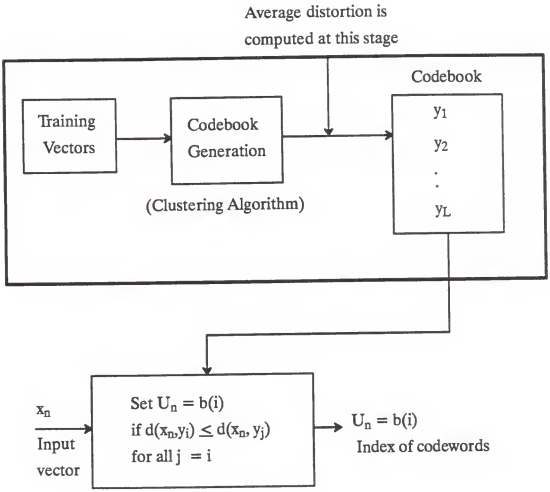


Figure 4–15. Schematic diagram of the vector quantizer.

one of a finite set of values, $S=\{y_i, 1 \le i \le L\}$. The set S, or the collection of possible reproduction vectors, is called the reproduction book or, simply, codebook of the quantizer, and L is the size of the codebook and its members are called codewords. The size L of the codebook is also called the number of levels, a term borrowed from scalar quantization terminology. To design such a codebook, we need to partition the N-dimensional space of the input vector into L regions or cells $\{C_i, 1 \le i \le L\}$ using a large number of training vectors. Thus, the process of codebook design is also known as training the codebook.

Typically, VQ works as follows. When the input vector $\mathbf{x}_n$ becomes available, the distortion (dissimilarity) between the input vector and each stored codeword is computed. The encoded output is then the binary representation of the index of the minimum distortion codeword. Since we represent the N-dimensional input vector with simply the index of the code vector, considerable data reduction is achieved.

The V/U/M/S and nasal/nonnasal classification systems based on the VQ codebook approach are shown in Figure 4−16 and Figure 4−17.

4.2.3.1.2 Distortion measures

A distortion measure d is an assignment of a nonnegative cost function d(x,y) to the process of reproducing any input vector **x** as a reproduction vector **y**. Given such a distortion measure, we can quantify the performance of a system by an average distortion between the input and the final reproduction. To be useful, a distortion measure should be tractable and computable to permit analysis, and subjectively meaningful so that large or small quantitative measures correlate with poor or good subjective quality. Therefore, the choice of the distortion measure is a key component in the VQ technique. There are many such distortion measures which are useful in speech analysis (Nocerino et al., 1985). We have adopted the Itakura-Saito distortion

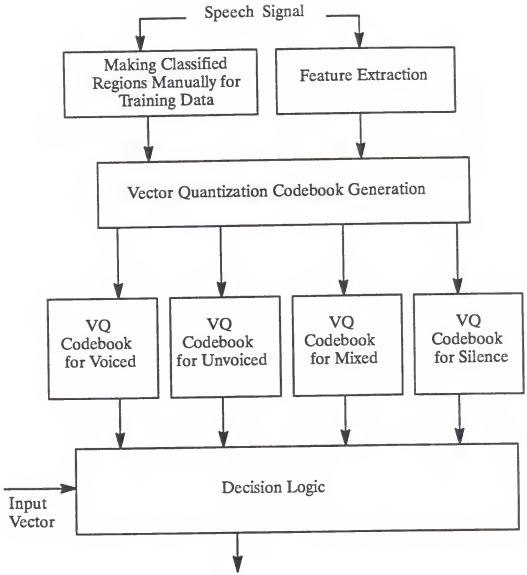


Figure 4−16. Block diagram of VQ for the V/ U/ M/ S decision.

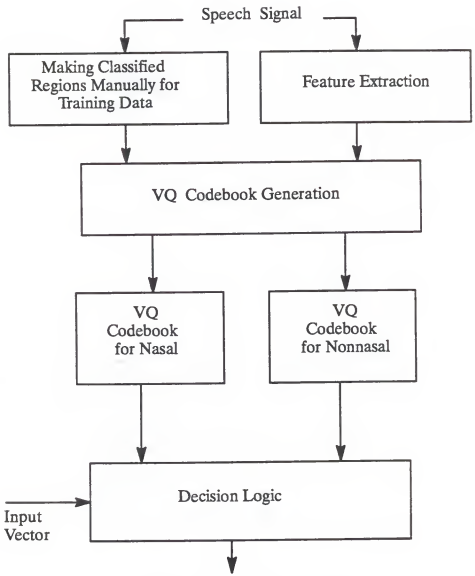


Figure 4–17. Block diagram of VQ for the nasal/nonnasal decision.

measure (Gray et al., 1980;Itakura and Saito, 1968;Linde et al., 1980), which in modified form is given by

$$d(x,y) = (x-y)^T R_x (x-y) \qquad (4-6)$$

where

$$R_x = \{ r(i-k)/r(0),\ 0 \geq i,\ k \leq N-1 \} \qquad (4-7)$$

is the autocorrelation matrix whose coefficients r(i-k) are used to compute the input vector **x** in eq. (4 – 6). This distortion resembles the form of the quadratic distortion measure, but the weighting matrix R_x depends on the input vector **x**, while the quadratic distortion measure does not.

4.2.3.1.3 Codebook design

In order to design an L-level codebook, as mentioned earlier, we need to partition the N-dimensional space into L cells $\{C_i,\ 1 \leq i \leq L\}$ using known training vectors. We associate each cell C_i with a vector $\mathbf{y_i}$. The quantizer then assigns the code vector $\mathbf{y_i}$ if the input vector **x** is in C_i. Each code vector $\mathbf{y_i}$ is chosen to minimize the average distortion in cell C_i. We call such a vector the centroid of the cell C_i. Computing the centroid for a particular region will depend upon the definition of the distortion measure. With the distortion measure of eq. (4 – 6), the centroid is given by

$$\mathbf{y_i} = \frac{\sum_j r_{ij}\ x_j}{\sum_j r_{ij}}, \text{ for } all\ i \qquad (4-8)$$

where r_{ij} represents the (i,j) element of the autocorrelation matrix with i representing the row and j the column; x_j represents the jth element of the input vector, **x**, of the feature parameters.

We used an iterative clustering algorithm to design the codebook, which requires an initial code vector. We used the technique of (Buzo et al., 1980) because it has been reported to give satisfactory results.

Using the feature vectors as training sequences, the codebook is generated as follows:

Step 1. Initialization.

Set n=0 and D_{avg} = 10000 (arbitrarily chosen as a large positive number). Fix $L=2^n$, n an integer and L as the largest number of levels desired.

Step 2. Find the centroid y_o of all the training vectors using eq. (4 − 8). This y_o is used as an initial code vector.

Step 3. Set n=n+1

Step 4. Partition the complete set of training vectors iteratively, until the decrement of D_{avg} at each iteration is less than a predetermined threshold, in such a way that the average distortion

$$D_{avg} = \frac{1}{M}\sum_{i=1}^{M} \min_{1 \le j \le 2^n} d(x_i, y_j) \qquad (4-9)$$

is minimized over the entire training set, i.e., M training vectors, where the distortion between vector x_i and y_j is denoted as $d(x_i, y_j)$. The vector y_j represents the centroid of each partition C_j, $1 \le j \le L$ and it becomes the codeword (Buzo et al., 1980).

Step 5. If the desired level of the codebook is met, i.e., $2^n \ge L$, then stop, otherwise go back to step 4.

In step 4, the decrement of D_{avg} to terminate the iteration at each level of the codebook is defined as

$$\text{DECREMENT} = \frac{\text{Previous } D_{avg} - \text{Current } D_{avg}}{\text{Current } D_{avg}} \qquad (4-10)$$

The threshold to terminate iteration was selected as 0.0001. The average distortion obtained using eq. (4 – 8) and the codebook generated in step 4 were examined for the classification of a speech signal.

4.2.3.1.4 Selection of codebook size

For this study the distortion measure for the VQ procedure is the Itakura-Saito distortion (Gray et al., 1980; Itakura and Saito, 1968; Linde et al., 1980). We conjecture that the average distortion for each subject might vary depending on the size of the codebook. The size of the codebook is to be chosen so that the maximum separation in distortion was to be obtained in V/U/M/N/S classification. Five sentences discussed in section 4.2.4 are used in this study. Each sentence was spoken by six speakers, three male and three female. Training for the VQ is performed using five sentences spoken by two male and two female. Using these training data we vary the codebook size from one to ten. We calculate the distortion for each codebook size and determined the classification errors for V/U/M/N/S classification using the training data. We found that increasing the codebook size beyond six did not reduce the number of classification errors. We therefore selecte our codebook size to be six. Thus, we quantize the average distortion measure to six "levels."

4.2.3.2 Neural Network

We use the feedforward neural network classifier, trained with the back-propagation algorithm, for the automatic V/U/M/N/S classification.

4.2.3.2.1 Multi-layer perceptron

Multi-layer perceptrons are feed-forward nets with one or more layers of nodes between the input and output nodes. These additional layers contain hidden units or nodes that are not directly connected to both the input and output nodes. Multi-layer perceptrons overcome many of the limitations of single-layer perceptrons. The capabilities of multi-layer perceptrons stem from the nonlinearities used within nodes. Similar behavior is exhibited by multi-layer perceptrons with multiple output nodes when sigmoidal nonlinearities are used and the decision rule is to select the class corresponding to the output node with the largest output. These nets can be trained with the back-propagation training algorithm (Rumelhart, 1986). An example of the behavior of a three-layer perceptron with N continuous valued inputs, M outputs and two layers of hidden units is presented in Figure 4−18.

The back-propagation algorithm is given as the logistic activation function

$$o_j(x) = \frac{1}{1 + e^{[-(W_j X + W_{0j})]}} \qquad (4-11)$$

where $o_j(x)$ is unit j's activation value when presented the input vector X, W is a column vector containing unit j's input weights, and W_{0j} is unit j's scalar bias weight (Rumelhart, 1986).

To train the network,we use the following back-propagation formula to modify a unit's weights

$$\Delta W_{ji} = \eta \delta_j o_i, \qquad (4-12)$$

in which ΔW_{ji} is the change to be made to the weight of unit j's input from unit i, η is the learning rate, δ_j is an error signal available at unit j, and o_i is the activation value of unit i after the input training vector has been presented to the network as input and

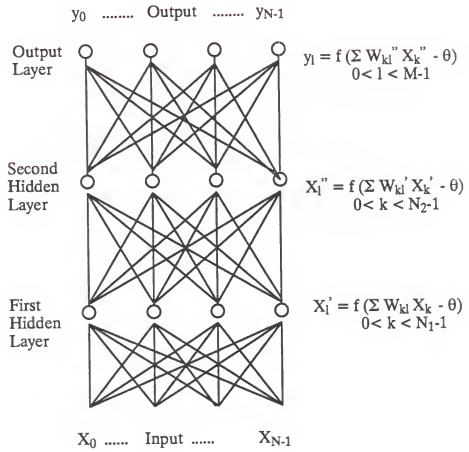


Figure 4–18. A three - layer perceptron with N continuous valued inputs.

the network has settled. The way of computing the error signal δ for each unit depends on the type of unit involved. If unit j is an output unit, the error signal is computed as

$$\delta_j = (t_j - o_j)\ o_j\ (1 - o_j), \qquad (4-13)$$

where t_j is the target vector element corresponding to output unit j. If unit j is a hidden unit, the error signal is computed as

$$\delta_j = o_j\ (1 - o_j)\ \Sigma\ \delta_x\ w_{kj} \qquad (4-14)$$

Here k indexes all units which have unit j as an input. δ_x is the error signal available at unit k, and w_{kj} is the weight of unit k's input from unit j. We use a feed-forward network, so all the units indexed by k are in layers succeeding the layer of unit j, and error signals propagate back from the output layer. we use a recursive algorithm starting at the output modes and working back to the first hidden layer. We adjust weights by

$$W_{ij}(t + 1) = W_{ij}(t) + \eta\delta_j X_i' \qquad (4-15)$$

Convergence is sometimes faster if a momentum term is added and weight changes are smoothed by

$$W_{ij}(t + 1) = W_{ij}(t) + \eta\delta_j X_i' + \alpha(W_{ij}(t) - W_{ij}(t - 1)), \qquad (4-16)$$

where $0 < \alpha < 1$.

4.2.3.2.2 Network structure

Eleven features for V/U/M/S and three features (SZCR, ratio7, ratio8) for nasal/nonnasal classification are used with different number of hidden units and layers and output units. Each input-unit output is connected to an input of every hidden unit and to an input of every output unit. Each hidden unit output was connected to an input

of every output unit. Connection weights are initially set to small real pseudo-random numbers. The number of hidden units in the network will be good near the high end of what seemed reasonable for the problem; a relatively large number of hidden units will be used to safeguard against training difficulties.

The output units are meant to be binary indicators of patterns, the first output indicating V, the second output indicating U, the third output indicating M, and the third output indicating S for the V/U/M/S classification (the first output indicating nasal and the second output indicating nonnasal for nasal/nonnasal classification). Decision logic is used to select the largest value of output units.

4.2.3.2.3 Data normalization

In general, a net is made up of sets of nodes arranged in layers. The outputs of nodes in one layer are transmitted to nodes in another layer through links that amplify or attenuate or inhibit such outputs through weighting factors. Except for the input layer nodes, the net input to each node is the sum of the weighted outputs of the modes. Each node is activated in accordance with the input to the node, the activation function of the node, and the bias of the node. Three activation functions, hard limiters and threshold logic elements and sigmoidal nonlinearities, is shown in Figure 4−19. The levels of these functions has between +1 and -1 or between +1 or -1 for sigmoid function. We will use the sigmoid function for the activation function, so the input data should be normalized between +1 and 0. The outputs of the nodes in that layer may be taken to be equal to the inputs, or we can take the opportunity to normalize those inputs in the sense that they can be scaled to fall between the values of -1 and +1. In particular, if

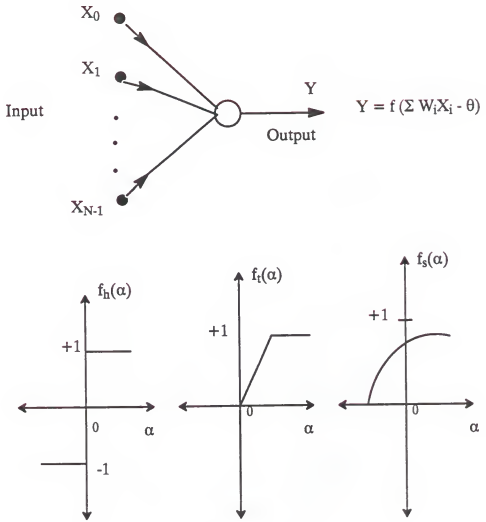


Figure 4–19. Computational element or node.

$$o_j(x) = \frac{1}{1 + e^{[-(W_j^t X + \theta_{0j})]}} \qquad (4-17)$$

then

$$\frac{d(o_j(x))}{d(net_j)} = o_j(1 - o_j) \qquad (4-18)$$

where

$$net_j = \mathbf{W}_j^t \mathbf{x} + \theta_{0j} \qquad (4-19)$$

Note that the thresholds θ are learned in the same manner as are the other weights. We simply imagine that θ is the weight from a unit that always has an output value of unity. Also note that the derivative $d(o_j) / d(net_j)$ reaches its maximum for o = 0.5 and, since $0 <= o_j <= 1$, approaches its minima as o approaches zero or one. Since the change in weight is proportional to this quantity, it is clear that weights that are connected to units in their midrange are changed the most. It is important to note that, for the activation function given by eq. (4 − 17), a node cannot have output values of 1 or 0 without infinitely large positive or negative weights. Therefore, in learning mode the values of 0.9 and 0.1 might suffice for specifying binary target output values.

4.2.3.2.4 Complexity requirements of the feedforward multilayer machine.

In a connectionist net that learns discriminants for pattern classification, it is natural to ask how many layers are required for the learning and discrimination task, and how many nodes are required in each layer. A simple argument provides insight (Lippman, 1987). Complications aside, we can argue that a three-layer machine can form arbitrarily complex decision regions and can therefore separate populations of

patterns even though such distributions might be intermeshed spatially in pattern space. Each first-layer node forms a hyperplane in pattern space, because the input to the node is the linear sum of the inputs. As the inputs change in value, the linear sum traces a line in two-dimensional space, a plane in three-dimensional space, and hyperplanes in N-dimensional space. This view indicates that each hyperregion requires 2N nodes in the first hidden layer, one node for each of the sides of the hyperregion. At the next layer, a node is needed to carry out an AND operation on that collection of hyperplanes. Thus, whereas a node in the first internal layer forms a hyperplane, a node in the second internal layer forms a hyperregion from the first layer nodes.

In practice, an excessive number of nodes in any one layer can generate noise. On the other hand, fault tolerance can be obtained with such redundancy in the number of nodes. Sometimes, we can simplify difficult learning tasks by increasing the number of internal layers. We can certainly do so when there are but three to four internal layers. However, Rumelhart, Hinton and Williams (1986) found that exponential increases in the number of hidden units were necessary to produce a linear increase in learning speed on the EXCLUSIVE-OR problem. Other researchers have found that increasing the number of hidden layers actually decreased the rate of learning in the random vector-pairing problem. The tendency is to try to expedite the learning process by increasing the extent of nonlinearity, rather than by increasing the complexity of the semilinear net.

We investigate that the requisite mapping can be carried out more directly on the input patterns . Therefore, in this research, we decide the complexity of net using the experimental results in various numbers of layers and nodes. According to classification results depending on the network complexity, the best architectures for net are following:

for the V/U/M/S classification

input nodes: 11

output nodes: 4

no. of hidden layer: 4

no. of units for hidden layers: 11, 11, 11, and 11

learning rate eta: 0.9

momentum rate alpha: 0.1

normalized system error: 0.0437

max. no. of iteration: 2000

for the nasal/nonnasal classification

input nodes: 3

output nodes: 2

no. of hidden layer: 3

no. of units for hidden layers: 8, 11, and 5

learning rate eta: 0.9

momentum rate alpha: 0.1

normalized system error: 0.0126

max. no. of iteration: 2000

4.2.3.2.5 Training and testing

In the learning phase of training such a net, we present the pattern as input and ask that the net adjust the set of weights in all the connecting links and also all the thresholds in the nodes such that the desired outputs are obtained at the output nodes. Once this adjustment has been accomplished by the net, we present another pair of input data and desired outputs and ask that net learn that association also. In fact, we ask that the net find a single set of weights and biases that will satisfy all the (input, output) pairs presented to it. Discrepancies between actual and target output values

again result in evaluation of weight changes. After complete presentation of all patterns in the training set, a new set of weights is obtained and new outputs are again evaluated in a feed forward manner. In a successful learning exercise, the system error will decrease with the number of iterations, and the procedure will converge to a stable set of weights, which will exhibit only small fluctuations in value as further learning is attempted.

Five sentences spoken by six speakers, three male and three female are used in this study. For the training data of V/U/M/S classification, a total of 200 frames with 50 frames for each subject are used. To obtain the optimal training condition, the same number of data set for each subject is recommended. For the training data of nasal/nonnasal classification, a total of 400 frames with 200 frames for each subject are used.

There are several other issues we need to keep in mind when we implement such nets.

Momentum, α and learning rate, η

There is the question of how the value of η is to be chosen. This is not a new or unusual problem; it is common to all steepest-descent methods of locating minima of functions. As might be expected, a large η corresponds to rapid learning but might also result in oscillations. According to Rumelhart, Hilton, and Williams(1986), we write

$$W_{ij}(t+1) = W_{ij}(t) + \eta\delta_j X_i' + \alpha(W_{ij}(t) - W_{ij}(t-1)), \qquad (4-20)$$

$$\text{where } 0 < \alpha < 1 .$$

where the quantity (t+1) is used to indicate the (n+1)th step, and α is a proportionality constant. The third term in eq. (4 – 20) is used to specify that the change in W at the (t+1)th step should be somewhat similar to the change undertaken at the n-th step. In this way, some inertia is built in, and momentum in the rate of change is conserved to some degree. Examination of the system error E over a large number of steps in the

iterative approach to a solution will generally show that a finite α tends to dampen the oscillations but can also serve to slow the rate of learning.

If weight corrections are carried out after presentation of each pattern, the method is not truly a gradient search procedure. In addition, the value of η needs to be small; otherwise, large excursions can take place in weight space. In this study, values of α and η is changed and evaluated.

Symmetry breaking

Learning procedure has one more problem that can be readily overcome. This is the problem of symmetry breaking. If all weights start out with equal values and if the solution requires that unequal weights be developed, the system can never learn. This is because error is propagated back through the weights in proportion to the values of the weights. This means that all hidden units connected directly to the output units will get identical error signals, and, since the weight changes depend on the error signals, the weights from those units to the output units must always be the same. We counteract this problem by starting the system with small random weights. Under these conditions symmetry problems of this kind do not arise.

Local minima

Single layer linear systems always have bowl-shaped error surfaces. However, in multilayer networks there is the possibility of rather more complex surfaces with many minima. Some of the minima constitute solutions to the problems in which the system reaches an errorless state. All such minima are global minima. However, it is possible for some of the minima to be deeper than others. In this case, a gradient descent method may not find the best possible solution to the problem at hand. Part of the study of back propagation networks and learning involves a study of how frequently and under what conditions local minima occur. In problems with many hidden units, local minima seem quite rare. However with few hidden units, local minima can be more common.

4.2.3.2.6 Criteria to stop the training

For the purpose of stopping the training, three conditions are checked to see whether learning should terminate. First, the maximum iteration number which was selected by the user in program is reached. Second, the error for each pattern selected is small enough. Third, system total error selected is small enough.

The derivation of the back propagation paradigm supposes that we are taking the derivative of the error function summed over all patterns. In this case, we might imagine that we would present all patterns and then sum the derivatives before changing the weights. Instead, we can compute the derivatives on each pattern and make the changes to the weights after each pattern rather than after each epoch.

Learning is carried out by the two commands. The first carries out training in sequential order, the second in permuted order. Training goes on until a maximum number of iteration is reached, or until the value of individual error becomes less than the value of a control parameter called maximum individual error for error criterion. For each, the square of the error is

$$E_p = 0.5 \Sigma (t_j - o_j)^2 \qquad (4-21)$$

and the average system error is

$$E = 1/2*P \Sigma\Sigma (t_j - o_j)^2 \qquad (4-22)$$

where the factor of one -half is inserted for mathematical convenience.

Figure 4−20 gives a flowchart of the feed-forward back-propagation algorithm for NN classifier.

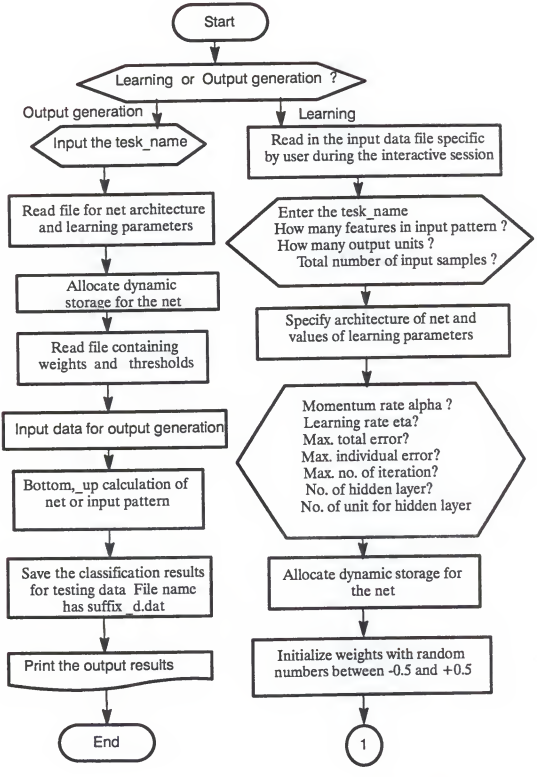


Figure 4−20. Flowchart of feed-forward back-propagation algorithm.

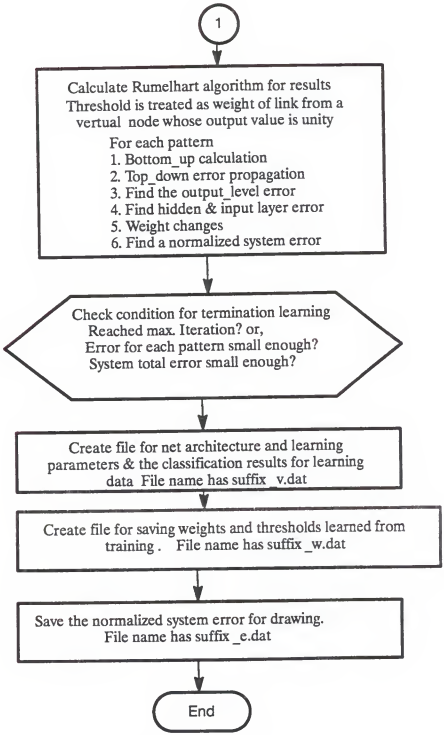


Figure 4–20. Continued

4.2.3.3 Decision Tree Method

A block diagram of the analysis and decision tree algorithm is shown in Figure 4−21. The speech signal is smoothed and is formatted into blocks of 100 samples (an interval of 10ms at 10kHz sampling frequency). For each block, the features are calculated and are used for an early classification of the frames that are clear cases of voiced and unvoiced. Statistics, such as averages and standard deviations, are calculated using the features of these clear-cut frames for use directly in the tree-structure pattern classification algorithm. In that step, the remaining more difficult input speech segments are assigned to all four categories of voiced, unvoiced, mixed, or silence according to a tree-structure pattern classification technique using the features and their statistics. For the voiced frames, the nasal/nonnasal decision and the detail mixed detection are following. The last step is the error correction step, where errors such as VVVUVVV and SSSUSSS are corrected to VVVVVVV and SSSSSSS.

4.2.3.3.1 V/U/M/S classification

In Figure 4−22, details of the tree-structure pattern classification algorithm for the one-channel four-way classifier are shown. The threshold values and decision rules in this figure were determined based on the five features (including the spectral ratios) and their statistics.

4.2.3.3.1.1 Feature extraction

It must be remembered in going through this step that in feature extraction statistics are only evaluated for the clear-cut voiced and unvoiced frames (and five beginning silent frames), rather than for a preclassified training sentence. The merit of this strategy, as opposed to the latter, is that our algorithm will have the capability of adaptation to the properties of any input sentence by changing its threshold values

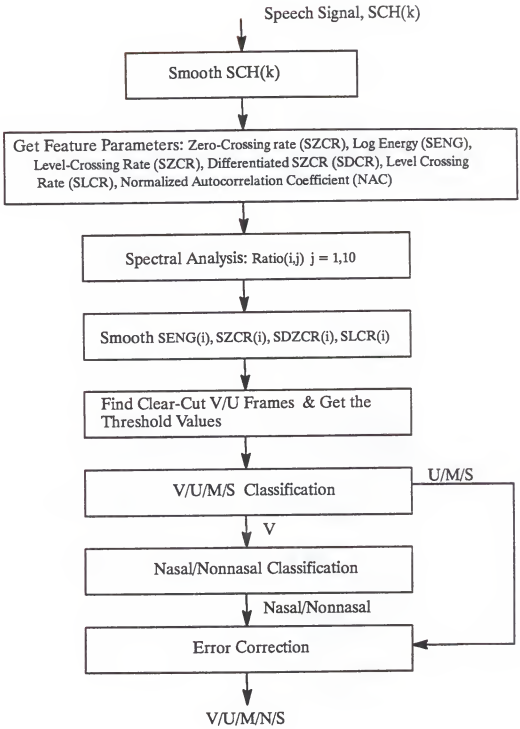


Figure 4−21. Block diagram of V/U/M/N/S classification using decision tree method.

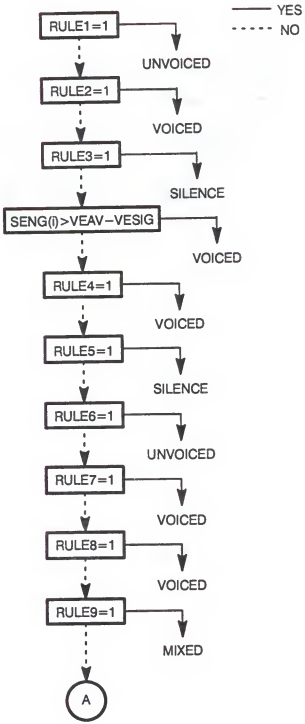


Figure 4−22. Details of decision tree classifier.

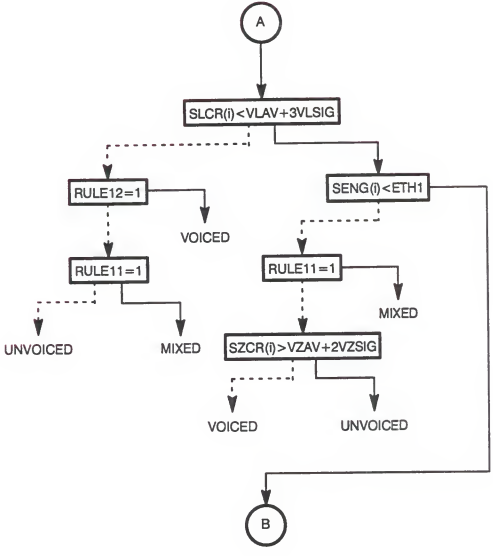


Figure 4−22. -- Continued

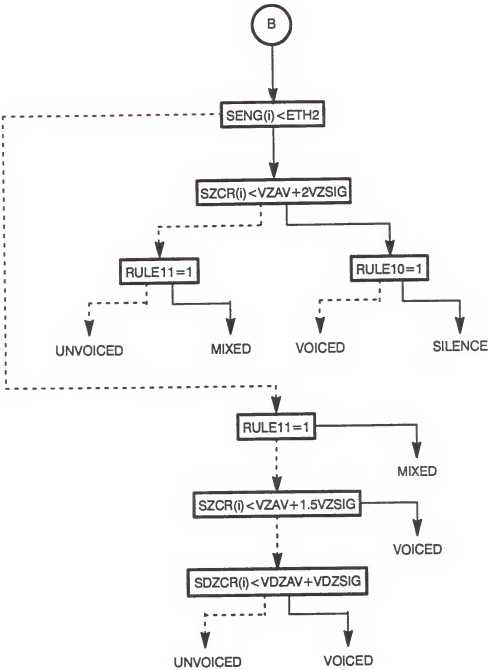


Figure 4—22. --Continued

automatically. (Most algorithms reviewed in Chapter 1 were not adaptive ones and could produce unacceptable results when a strange speaker was subjected.)

In order to understand the details of this feature extraction step, explanation of the parameters is essential. In the following definitions the index 'i' was used for frames, the index 'k' was used for data points across frames, while 'j' was used for data points within a frame.

SCH(k): the k-th data point of the speech signal.

DSCH(k): the k-th data point of the differentiated speech signal.

$$DSCH(k) = SCH(k) - SCH(k-1) \qquad (4-23)$$

SENG(i): the energy in decibel (dB) of the i-th frame of the speech signal.

SZCR(i): the zero crossing rate of the i-th frame of the speech signal. The value of SZCR(i) is incremented by one when the product of SCH(i*100+j) and SCH(i*100+j-1) is less than zero.

SDZCR(i): the zero crossing rate of the differentiated speech signal.

SLCR(i): the level crossing rate of the i-th frame of the speech signal. The value of SLCR(i) is incremented by one when the product of (SCH(i+100+j)-SLTT) and (SCH(i*100+j-1)-SLTT) is less than zero.

IVUS(i): the classification result of the i-th frame represented in number. The values 1, 3, 7, and 9 are arbitrarily assigned to silent, unvoiced, mixed, and voiced, respectively.

SLTT: The threshold value to calculate the level crossing rate of the speech signal. It was set to 10% of the average magnitude of rectified voiced/mixed sounds.

VEAV: the average energy of voiced sounds.

VESIG: the standard deviation of the energy of voiced sounds.

VZAV: the average zero crossing rate of voiced sounds.

VZSIG: the standard deviation of the zero crossing rate of voiced sounds.

VLAV: the average level crossing rate of voiced sounds.

VLSIG: the standard deviation of the level crossing rate for voiced sounds.

VDZAV: the average zero crossing rate of differentiated voiced sounds.

VDZSIG: the standard deviation of the zero crossing rate for differentiated voiced sounds.

The eight statistics above (with variable names in "V") are all calculated based on the frames classified as clear-cut voiced frames in the "early" classification. There are analogous statistics for the clear-cut unvoiced frames (with variable names in "U"), and for the five silent frames at the beginning of each utterance (with variable names in "S"). All these statistical values are calculated on a sentence by sentence basis, using the clear-cut voiced and unvoiced frames and the five silent frames in the sentence, and can therefore be considered as adaptive statistical values.

In smoothing SCH(k), SENG(i), SZCR(i), SDZCR(i), and SLCR(i), a three-point filter of (0.12, 0.76, 0.12) was used. For example:

$$SCH(k) = 0.12*\{SCH(k-1)+SCH(k+1)\}+0.76*SCH(k) \qquad (4-24)$$

This filter has linear phase characteristics and plays a similar role to a low pass filter.

All these features and statistics are used to determine threshold values for the second step, pattern classification.

4.2.3.3.1.2 Threshold explanation

The averages and standard deviations for these features are calculated for the "clear-cut" voiced, the "clear-cut" unvoiced, and the five beginning silent frames. The early classification of these "clear-cut" frames are achieved by applying simple rules. A frame is labeled as unvoiced, if all five spectral ratios (as defined in the next section) are less than zero. When all five ratios are greater than 20, the frame is classified as voiced. If there is no clear-cut unvoiced frame in a given sentence, the threshold values were defined as

$$ETH1 = (SEAV+VEAV)/2 \quad (4-25)$$

$$ETH2 = (SEAV+VEAV-2*(SESIG+VESIG))/2 \quad (4-26)$$

$$ETH3 = (SEAV+VEAV+0.5*(SESIG+VESIG))/2 \quad (4-27)$$

When there were some clear-cut unvoiced frames, these values were set as follows.

$$ETH1 = UEAV \quad (4-28)$$

$$ETH2 = UEAV-UESIG \quad (4-29)$$

$$ETH3 = UEAV+UESIG \quad (4-30)$$

4.2.3.3.1.3 Decision rules

The rules in our one-channel four-way pattern classification algorithm were defined as

RULE1 = 1, if all five ratios are less than zero

= 0, otherwise

RULE2 = 1, if all five ratios are greater than 30

= 0, otherwise

RULE3 = 1, if SENG(i) is less than SEAV

= 0, otherwise

RULE4 = 1, if SENG(i) is greater than VEAV-2*VESIG, SZCR(i) is less than VZAV+2*VZSIG, SLCR(i) is less than VLAV+2*VLSIG, and SDZCR(i) is less than VDZAV+ 2*VDZSIG

= 0, otherwise

RULE5 = 1, if SENG(i) is less than SEAV+2*SESIG, SZCR(i) is greater than SZAV-2*SZSIG, SZCR(i) is less SZAV+ 2*SZSIG, SDZCR(i) is greater than SDZAV-2*SDZSIG, SDZCR(i) is less than SDZAV+2*SDZSIG, and RATIO(i,5) is greater than zero.

= 0, otherwise

RULE6 = 1, if SENG(i) is less than UEAV+UESIG, SENG(i) is greater than UEAV-1.5*UESIG, SZCR(i) is less than UZAV+UZSIG, SZCR(i) is greater than UZAV- 1.5*UZSIG, SLCR(i) is less than ULAV+ULSIG, and SLCR(i) is greater than ULAV-1.5*ULSIG.

= 0, otherwise

RULE7 = 1, if all five ratios are greater than 20

= 0, otherwise

RULE8 = 1, if all three of RATIO(i,1), RATIO(i,2), and RATIO(i,3) are greater than 30, RATIO(i,4)+RATIO(i,5) is greater than zero, and SENG(i) is greater than ETH1

= 0, otherwise

RULE9 = 1, if the sum of RATIO(i,4) and RATIO(i,5) is less than -10, the sum of RATIO(i,2) and RATIO(i,3) is greater than 20, and SENG(i) is greater than ETH3.

= 0, otherwise

RULE10 = 1, if SENG(i) is less than SEAV+5*SESIG and SLCR(i) is less than 3

= 0, otherwise

RULE11 = 1, if RATIO(i,1) is greater than zero, RATIO(i,2) or RATIO(i,3) is greater than zero, RATIO(i,4) or RATIO(i,5) is less than zero, SDZCR(i) is greater than VDZAV+1.25*VDZSIG, STH is greater than 8, SDZCR(i) is greater than 40, and SENG(i) is greater than ETH2.

= 0, otherwise

RULE12 = 1, if SDZCR(i) is less than VDZAV+VDZSIG and SENG(i) is greater than VEAV-1.5*VESIG

= 0, otherwise

The basic criterion for distinguishing mixed sounds from voiced sounds is that the energy of mixed sounds is almost equally distributed over the entire frequency range, while the energy of voiced sounds is concentrated in the low frequency range. Hence, if the situation is ideal, we will get large values for all five ratios for voiced sound while the values for mixed sounds are low, close to one. In practice, it is observed that mixed sounds were affected a great deal by their neighboring sounds. This phenomenon makes it extremely hard to identify mixed sounds located within a transition interval. Hence, the normalized autocorrelation coefficient is used to detect the mixed sound in the voiced regions.

The normalized autocorrelation coefficient is the correlation between adjacent speech samples, and, by definition, varies between -1 and +1. Due to the concentration of low-frequency energy in voiced sounds, adjacent samples of voiced speech waveform are highly correlated and the parameter is close to unity (is close to zero for unvoiced speech). On the other hand, the correlation coefficient for the mixed sounds is mostly in the areas less than 0.7 shown in Table 4−1.

4.2.3.3.1.4 Error correction

In Figure 4−23, the procedure of the error correction step for the smoothing of the results is presented. In single voiced frame error correction, a mixed frame is corrected to voiced if one of its neighboring frames is voiced and the other is silent. This correction is based on the observation that a mixed sound which begins or ends an utterance would last more than 20 milliseconds. (Some mixed sounds only last one frame in voiced-to-unvoiced or unvoiced-to- voiced transition regions, but this case is excluded from this correction step.)

The single mixed frame error correction step tries to correct some mixed-to-voiced or mixed-to-unvoiced misclassifications. If the sum of RATIO(i,4) and RATIO(i,5) is less than zero and SDZCR is greater than VDZAV+VDZSIG, then

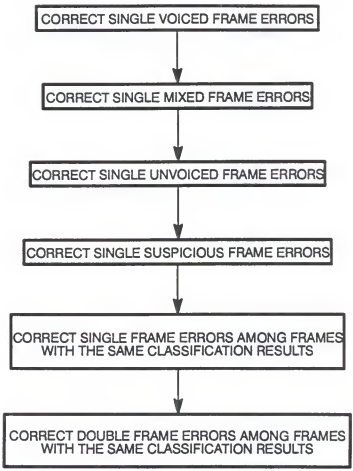


Figure 4−23. Error correction step.

the current voiced frame is corrected to mixed. When the sum of the RATIO(i,1), RATIO(i,2), and RATIO(i,3) is greater than 18, the sum of RATIO(i,4) and RATIO(i,5) is less than 15 but positive, and SENG is greater than ETH1, then the current unvoiced frame is reclassified as mixed. The single unvoiced frame error correction step corrects a voiced frame to unvoiced, when both RATIO(i,4) and RATIO(i,5) are less than zero.

Next is the single suspicious frame error correction step. A single voiced or unvoiced frame is corrected to the category of the following frame based on a simple distance measure, i.e., the taxicab distance measure of the zero crossing rate of the speech signal. Specifically, if |SZCR(i-1)-SZCR(i)| is greater than |SZCR(i+1)-SZCR(i)| and all three frames are classified into different categories, then the current frame is reclassified to the category of the following frame. The last two steps, single and double frame error correction are performed in the same way here as they are in the two-channel classifier.

4.2.3.3.2 Nasal/nonnasal classification

Based on the observation of the acoustic characteristics of nasals in Table 4−2, we develop the following algorithm:

1) Select the voiced frames from the result of the V/U/M/S classification.

2) Find the nasal candidates by comparing the log energy with the average energy of voiced sounds, VEAV. If SENG(i) > 0.95*VEAV, then the frame is a nasal candidate.

3) If ZCR(i) <7.2, ratio7 > 3.3, ratio8 > 21, and ratio10 < 22, then the segment is a nasal consonant.

4) If ZCR(i) <6.9, ratio7 > 2.8, ratio8 > 21, and ratio10 < 22, then the segment is a nasal consonant.

4) If ZCR(i) < 8, ratio7 > 1.1, ratio9 <42, and ratio10 < 12, then the segment is a nasal consonant.

4.2.4 Result

For the data base of the V/U/M/N/S classification algorithm in this study, five sentences are selected based on their phonetic contents. Each sentence was spoken by six speakers, three male and three female. These five sentences are as following:

1) We were away a year ago. (Voiced.)

2) Early one morning a man and a woman ambled along a one mile lane. (Voiced and nasals.)

3) Should we chase those cowboys? (Fricatives and plosives.)

4) That zany van is azure. (Voiced fricatives, i.e., mixed.)

5) We saw the ten pink fish. (Unvoiced plosives and fricatives.)

This data is more extensive than that used in (Atal et al., 1976; Rabiner et al., 1977; Siegel et al., 1982; Rabiner et al., 1977).

Figure 4−24 show the results of the V/U/M/S classification for sentence 4 spoken by a male speaker. The first plot is the spectrogram of the speech signal. The second one is the result by the manual procedure. The third and fourth plots are the results of the V/U/M/S classification by the two-channel algorithm (Childers et al., 1989) and by the one-channel algorithm respectively. These contours can assume one of four values where value 1 is silence, value 3 is unvoiced, value 7 is mixed, and value 9 is voiced. It can be seen that the two-channel algorithm made essentially two error parts in classifying mixed intervals as voiced. The one-channel algorithm corrected one error part and left the rest of the V/U/M/S contour the same.

Figure 4−25 illustrates the V/U/M/N/S classification combined with the V/U/M/S and the nasal/nonnasal classifications for sentence 4 spoken by a male speaker. The first plot is the spectrogram of the speech signal. The second and third plots are the V/U/M/N/S classification by the manual procedure and by the algorithm respectively. These contours can assume one of four values where value 1 is silence,

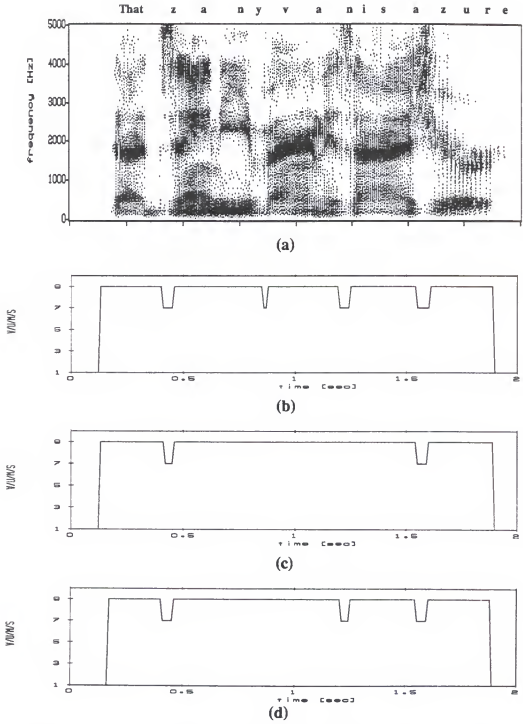


Figure 4−24. Comparison of V/U/M/S classification by the algorithm and manual procedures: (a) spectrogram, (b) manual classification, (c) two-channel classification algorithm, and (d) one-channel classification algorithm.

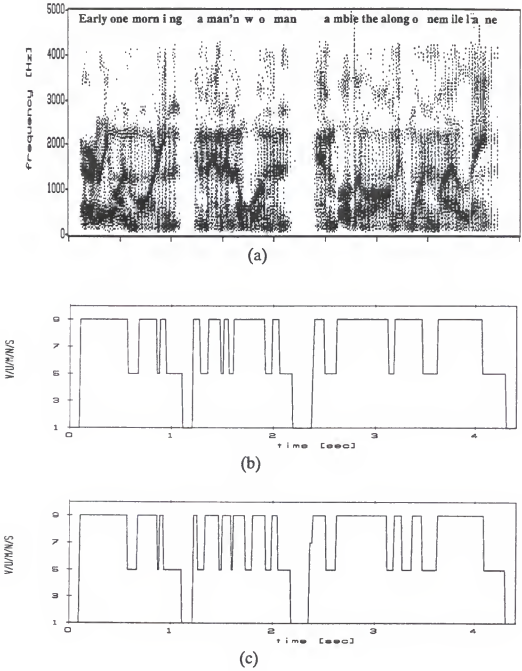


Figure 4−25. Comparison of V/U/M/N/S classification by the algorithm and the manual procedures: (a) spectrogram, (b) manual classification, and (c) one-CH five-way classification algorithm.

value 3 is unvoiced, value 5 is nasal, value 7 is mixed, and value 9 is voiced. This figure shows that two areas of voiced is classified as nasals and that the starting and ending of nasals make errors in classification. Although misclassification in several frames happened, the algorithm works fairly well. From Table 4−3 to Table 4−12, the performance of the classifiers used in training and in testing data sets is shown.

In V/U/M/S classification, training for the VQ is performed using sentence 1-5 spoken by two male and two female speakers. The training information consist of a total of 5685 frames. Testing consist of a total 2384 frames and was performed using the same sentences spoken by the two speakers (one male and one female) not used in training. Training for the NN consist of a total of 200 frames, 50 frames for each subject. Testing consist of a total 4181 frames not used in training.

In nasal/nonnasal decision, training for the VQ is performed using sentence 1-5 spoken by two male and two female speakers. Training consisted of a total of 4181 frames which are not classified as the silence, unvoiced, and mixed subjects in V/U/M/S classification. Testing consist of a total of 1649 frames classified as a voiced subject in V/U/M/S classification and is not used in training.

Clearly, it is desirable to use as few features as possible to perform the V/U/M/N/S classification, in order to minimize the computation time in analyzing speech. Moreover, when using statistical training methods, an insufficient number of known training samples can lead to what is known as the "dimensionality problem" (Wu at el., 1975) − the phenomenon that the performance of a classifier may degrade as the number of features used is increased (Papoulis, 1965). As a result of the feature selection procedure, the following features were used in the classifier:

V/U/M/S decision: SENG, C, SZCR, SDZCR, SLCR, Ratio1, Ratio2, Ratio3, Ratio4, Ratio5, Ratio6

Nasal/nonnasal decision: SZCR, Ratio7, Ratio8, Ratio9, Ratio10

For the NN and the VQ classifiers in the nasal/nonnasal decision, we only used SZCR,

ratio7 and ratio8 for the features. These features result in the improved classification on the training set.

Table 4−3 shows the V/U/M/S performance of the VQ classifier on the frames used in training. The entry in row and column of the table indicates how many classes of raw frame were classified as being in column class (e.g., from Silence, 2 Silent frames were classified as voiced). As can be seen from the table, overall classification accuracy of 97.5 % was obtained. Table 4−4 shows the V/U/M/S performance of the VQ classifier on the frames of testing not used in training. The performance has degraded somewhat (90.85 % overall accuracy, 80.55 % percent correct classification of mixed frames) that were available for training. Of the testing frames classified as mixed, only 80.55% (100% for training data) actually were mixed. This behavior may be partially explained by the less mixed frames in the speech used in training and in testing. Table 4−5 and Table 4−6 show the performance of the nasal/nonnasal classification of the VQ classifier for training and testing. The overall performance for training is 87.2 % and that for testing is 84.41 %.

Table 4−7 shows the performance of V/U/M/S classification of the NN classifier on the frames used in training. The number of frames for training is different from that of VQ (a total of 200 frames for the NN classifier and a total of 5685 frames for the VQ classifier). The overall performance is 97.5 % which is almost the same result in Table 4−3. Table 4−8 shows the performance for testing. An overall classification accuracy of 96.86% is obtained. Table 4−9 and Table 4−10 show the performance of nasal/nonnasal classification of the NN classifier. The number of frames used in training is 100 (4180 frames in VQ). As can be seen from the table, the overall classification accuracy is 94% for training and is 82.9 % for the testing, which are slightly better than those of the VQ classifier. As on the testing frames, the overall performance is better than that for the VQ classifier. It is worth noting that, although

we used less frames for training of the NN classifier, the NN classifier minimized overall misclassification.

Table 4−11 and Table 4−12 show the performance of the decision tree classifier. The overall performance of V/U/M/S classification is 97.06 % and that of nasal/nonnasal is 82.36 %. The performance has degraded somewhat, but the classification is still fairly accurate. As mentioned before, the decision tree method allows a more flexible division of a feature space and does not need the training procedure like the VQ and the NN classifiers.

4.3 Pitch Synchronous V/U/M/N/S Classification

Source-tract interaction has been established as an important factor for high quality speech synthesis. We can assess such interaction from the speech signal by using the pitch synchronous analysis which reduces the errors in the estimation of vocal tract parameters (Childers et al., 1985).

Several studies of analysis frame size and location are available (Chandra and Lin, 1974; Rabiner et al., 1976). Their conclusions are that the pitch-synchronous analysis method give significantly lower normalized squared prediction error and more accurate formant frequencies and bandwidths than the pitch asynchronous method. One of the factors that degrade the naturalness of synthetic speech is the spectral estimation of a speech signal over fixed frames. Wong(Wong, 1980) reported similar results, suggesting this to be the cause of a "warble" effect in the synthetic signal. Further, the gain values obtained from the analysis also show large fluctuations from frame to frame which further deteriorate the synthesis. The earlier results of Pinson (Pinson, 1963) using a least squares fit approach gave better estimates of formant frequencies and bandwidths, when the analysis was pitch synchronous.

Table 4–3. Performance of V/U/M/S classification using the VQ classifier on the frames used in training, compared to the manual classification.

Actual class ->	Silence	Unvoiced	Voiced	Mixed
Identified as Silence	1029	10	37	
Identified as Unvoiced	2	367	36	
Identified as Voiced	6	9	4052	
Identified as Mixed	1	3	68	95 / 95
Total 5543/5685 97.5 %	1029/1038 99.13 %	367/389 94.34 %	4052/4193 96.63 %	95/95 100 %

Table 4–4. Performance of V/U/M/S classification using the VQ classifier on the frames in test sentences, compared to the manual classification.

Actual class ->	Silence	Unvoiced	Voiced	Mixed
Identified as Silence	394	7	40	1
Identified as Unvoiced	8	181	33	2
Identified as Voiced	3	16	1562	4
Identified as Mixed	10	18	76	29
Total 2166/2384 90.85 %	394/415 94.93 %	181/222 81.53 %	1562/1711 91.29 %	29/36 80.55 %

Table 4–5. Performance of nasal/nonnasal classification using the VQ classifier on the frames in training sentences, compared to the manual classification.

Actual Class ->	Non - Nasal	Nasal
Identified as Non - Nasal	2993	65
Identified as Nasal	470	653
Total 3646/4181 87.2 %	2993/3463 86.42 %	653/718 91.0 %

Table 4–6. Performance of nasal/nonnasal classification using the VQ classifier on the frames in testing sentences, compared to the manual classification.

Actual Class ->	Non - Nasal	Nasal
Identified as Non - Nasal	1215	25
Identified as Nasal	232	177
Total 1392/1649 84.41 %	1215/1447 83.96 %	177/202 87.62 %

Table 4–7. Performance of V/U/M/S classification using the NN classifier on the frames in training data, compared to the manual classification.

Actual class ->	Silence	Unvoiced	Voiced	Mixed
Identified as Silence	50			1
Identified as Unvoiced		49		3
Identified as Voiced			50	
Identified as Mixed		1		46
Total 195/200 97.5 %	50/50 100 %	49/50 98 %	50/50 100 %	46/50 92 %

Table 4–8. Performance of V/U/M/S classification using the NN classifier on the frames in test sentences, compared to the manual classification.

Actual class ->	Silence	Unvoiced	Voiced	Mixed
Identified as Silence	1375	17	25	1
Identified as Unvoiced	5	493	9	5
Identified as Voiced	20	5	5717	9
Identified as Mixed	3	46	103	66
Total 7651/7899 96.86 %	1375/1403 98.00 %	493/561 87.87 %	5717/5854 97.65 %	66/81 81.48 %

Table 4–9. Performance of nasal/nonnasal classification using the NN classifier on the frames in training sentences, compared to the manual classification.

Actual Class ->	Non - Nasal	Nasal
Identified as Non - Nasal	187	11
Identified as Nasal	13	189
Total 376/400 94.00 %	187/200 93.5 %	189/200 94.5 %

Table 4–10. Performance of nasal/nonnasal classification using the NN classifier on the frames in testing sentences, compared to the manual classification.

Actual Class ->	Non - Nasal	Nasal
Identified as Non - Nasal	3862	80
Identified as Nasal	848	640
Total 4502/5430 82.90 %	3862/4710 81.99 %	640/720 88.88 %

Table 4–11. Performance of V/U/M/S classification using the statistical decision tree classifier on the frames, compared to the manual classification.

Actual class ->	Silence	Unvoiced	Voiced	Mixed
Identified as Silence	1421	25	55	
Identified as Unvoiced	22	517	25	15
Identified as Voiced	9	48	5814	13
Identified as Mixed	1	21	10	103
Total 7855/8099 96.98 %	1421/1453 97.79 %	517/611 84.61 %	5814/5904 98.47 %	103/131 83.62 %

Table 4–12. Performance of nasal/nonnasal classification using the statistical decision tree classifier on the frames, compared to the manual classification.

Actual Class ->	Non - Nasal	Nasal
Identified as Non - Nasal	3977	95
Identified as Nasal	933	825
Total 4802/5830 82.36 %	3977/4910 80.99 %	825/920 89.67 %

Pitch synchronous analysis performed over either one period-long frames or closed glottis regions produce the best result for analysis/synthesis (Childers and Wu, 1990). In the first step of the analysis the speech time waveform is segmented into pitch periods which are then subjected to a V/U/M/N/S classification. Utilizing the assumption that each period is one of an infinite sequence of identical periods, this "pitch synchronous" representation can be related easily and precisely to vocal characteristics.

In pitch synchronous V/U/M/N/S classification, two of the basic problems are the following:

1) Determination of the exact beginning and end of each pitch period during voiced speech segments. The choice of the exact beginning and ending of the pitch period is often quite arbitrary. For example, based on the acoustic waveform alone, some candidates for defining the beginning and end of the period include the maximum value during the period, the zero-crossing prior to the maximum, etc. The only requirement on such a measurement is that it be consistent from period-to-period in order to be able to define the "exact" location of the beginning and end of each pitch period. The lack of such consistency can lead to spurious pitch period estimates.

2) Selection of the features based on the pitch-period long frame which are variable.

For determining the exact beginning and ending of the pitch period, the algorithms for pitch determination are used based on the three methods discussed in section 4.1. Once the pitch period information is determined, the pitch-synchronous V/U/M/N/S algorithm is extended to make analysis interval synchronous with the regions of the glottis. The parameters and thresholds which are used in fixed frame size analysis are normalized in each pitch interval.

Figure 4–26 shows the procedure of the pitch synchronous V/U/M/N/S classification. In the first step of the analysis the speech signal is segmented into pitch

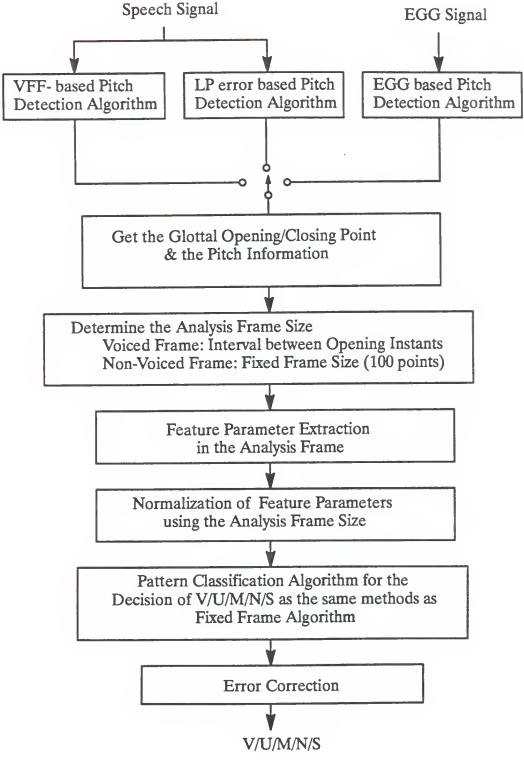


Figure 4−26. Overview of the pitch synchronous V/U/M/N/S algorithm.

periods using the pitch detection algorithms. To get the analysis frame for the non-voiced area, the pre-processing analysis for the decision of analysis interval is needed. Non-voiced areas can be divided into a 100 frame length for analysis. After deciding the analysis intervals for all the frames in the input signal, we extract the feature parameters as the same parameter used in the fixed frame size V/U/M/N/S algorithm and normalize the parameters into the value for one point by dividing by the analysis interval. Using the normalized feature parameters, the V/U/M/N/S procedure can be done with pattern recognition algorithms. Most cases of discrepancy in the speech segmentation using pitch-asynchronous method occur at the transition of classes. The pitch-synchronous V/U/M/N/S classification give the fine segmentation in the transition regions. The performance is almost the same as that in the pitch asynchronous methods discussed in section 4.2.

4.4 Summary

New time domain pitch detection algorithms, which provide pitch estimates on a period by period basis, have been described. The pitch smearing effect inherent in speech signal based methods is avoided. In comparison with the SIFT algorithm, it is shown that new algorithms have the performance which are closest to that obtained using the EGG for the speech of normal and pathologic speakers.

A fairly general framework based on a pattern recognition approach to V/U/M/N/S classification has been described in which a set of measurements are made on the interval being classified, and VQ, NN, and decision tree classifiers are used to select the appropriate class. The work constitutes a demonstration that the V/U/M/N/S classification can be made with reasonable accuracy. A VQ classifier achieve 97.5% classification accuracy on training and 90.85 % accuracy on testing in the V/U/M/S classification and achieve 87.02% on training and 84.41% for testing in nasal/nonnasal

classification. A NN classifier achieve 97.5 % accuracy on training and 96.86 % on testing in the V/U/M/S classification and achieve 94% on training and 82.9% for testing in the nasal/nonnasal classification. A decision tree classifier achieved 97.06 % in V/U/M/S classification and achieved 82.36 % in nasal/nonnasal classification. In summary, several pattern classification approaches to making the V/U/M/N/S decision is shown to produce good results.

For the nasal/nonnasal classification, the misclassification of nonnasal to nasal occurs mostly in the areas of the nasalized vowels. So our nasal/nonnasal algorithm needs to be combined with the spectral based algorithm by Yea (Yea and Childers, 1983) for the identification of the nasalized vowels.

The work also reveals some of the difficulties of the manual classification. There is necessarily some inaccuracy introduced by the manual classification; it is not always obvious what the correct classification for a frame is. For example, the classification rates of the NN and the VQ classifiers for the training data are 97.5%. The misclassification is probably caused by the inaccurate manual classification.

We have developed a pitch-synchronous V/U/M/N/S classification method. The method provides the information that is useful for fine segmentation of the speech signal and also provides a classification rate as good as the pitch-asynchronous algorithm discussed in this study. Pitch synchronous analysis method will be used in a complete analysis/synthesis system in the next chapter.

CHAPTER 5
ANALYSIS II: ADAPTIVE FORMANT TRACKING AND GLOTTAL INVERSE FILTERING (GIF) USING CLOSED PHASE WRLS-VFF-VT

5.1 Introduction

In this chapter we present two applications using the WRLS-VFF-VT algorithm for speech signal analysis. For a nonstationary speech signal, accurately tracking speech parameters like vocal tract resonance frequencies (formants) and their bandwidths are essential for the development of speech recognition and speech synthesis systems. Using the frame-based linear predictive coding (LPC) analysis, the accuracy of formant tracking of a speech signal is affected by 1) the position of the analysis frame, 2) the length of the analysis window, and 3) the time-varying characteristics of the speech signal. An adaptive filter approach, which tracks the time-varying parameters of the vocal tract and updates the parameters during the glottal closed phase interval, can reduce the formant estimation error.

In the first section of this chapter, we introduce a closed phase WRLS-VFF-VT algorithm for formant tracking and compare its performance with other methods such as closed phase covariance analysis.

In the second section, we discuss the estimation of a glottal excitation source using the glottal inverse filtering (GIF) analysis method. The WRLS-VFF-VT algorithm, which sequentially estimates the filter coefficients, estimation error, the variable forgetting factor, the pitch period, and the instant at which the glottis closes, can be easily applied to GIF analysis. For the estimation of the pitch period and the

glottal opening/closed instants, we can use the VFF signal, LP residual error signal and DEGG signal separately. The performance of this method is compared with other methods (i.e., the two-pass and two-channel methods).

5.2 Estimation of the Speech Parameters Based on the WRLS-VFF-VT

From our experience with a simplified cascade-parallel model of the vocal tract it appeared that the extraction of synthesizer control parameters from the LPC analysis of natural speech was quite complicated. Only in the vowels could the LPC formants be mapped onto the parameters of the cascade branch. In all other sounds, the approximation of the natural speech spectrum by selecting appropriate values of the parameters of the cascade and parallel branches appeared to be an art rather than a science. It was felt that a pole-zero model of the short-time speech spectrum would provide a much better fit to the acoustic behavior of the actual speech production apparatus, thus enabling us to interpret poles resulting from the analysis as genuine formants and to model anti-resonances in all speech sounds where they may occur. Using a cascade pole-zero model for synthesis would make large amounts of articulatory based phonetic knowledge available for guiding rule development.

Over the years there has been considerable interest in the time-varying modeling of speech signals (Liporace, 1975; Casacuberta and Vidal, 1987). The motivation behind this effort stems from the fact that many elements of speech such as stops, fricative onsets, and transitions between consonants and vowels, exhibit rapid changes that cannot be modeled satisfactorily by standard time-invariant techniques. In other words, an invariant model applied to relatively short segments of data is not appropriate in these situations. As a result, more meaningful feature sets for such events could be obtained by means of parametric models based on time-dependent difference equations which have been proposed in the recent literature (Grenier, 1983;

Hall et al., 1983). However, most methods employ least squares techniques for parameter extraction and these do not perform very well in the presence of additive noise or when the magnitudes of some of the coefficients are very small, both of which are distinct possibilities when dealing with real speech data. Furthermore, these models fail to differentiate regions of glottal excitation from those where the glottis is closed.

In order to estimate the vocal tract parameters accurately, it is necessary to eliminate the influence of the pitch period from the spectra of speech signals. If input pulses are estimated from the speech waveform and this estimated input is used for the model input, we can expect to obtain correct formants without the influence of pitch, that is, the correct parameters of the speech production model. In Chapter 3, we proposed a method of estimating ARMA parameters, input pulse train, and input white noise at the same time so that formants and antiformants of speech can be correctly estimated.

A comparison of different LP methods for speech analysis has been studied in (Krishnamurthy and Childers, 1986). The results showed the superiority of the two-channel pitch-synchronous closed-phase covariance (CPC) method in terms of estimation accuracy and tracking ability over several other common LP methods, such as LPC asynchronous, pitch-synchronous covariance, and pitch-synchronous circulation methods.

In this section the proposed methods are experimentally compared with the two-channel pitch-synchronous CPC method (Krishnamurthy and Childers, 1986). The experimental comparisons are performed for synthetic waveforms and real speech. We have chosen the formant frequencies and bandwidth estimated by LP analysis of speech to characterize the vocal tract filter.

5.2.1 Adaptive Formant Tracking using Closed Phase WRLS-VFF-VT

Past use of a pitch-synchronous closed phase analysis method has been limited due to difficulties associated with the task of accurately isolating the closed phase region in successive periods of speech. In this study, three pitch detection algorithms discussed in Chapter 4 can be used to isolate individual pitch periods of speech and to determine the closed glottal segments in each period.

Since the pitch synchronous closed phase method is known to give more accurate estimates of the vocal tract parameters than the pitch asynchronous method, we have implemented the following pitch synchronous closed phase WRLS-VFF-VT algorithm for speech analysis.

5.2.2 Closed Phase WRLS-VFF-VT Algorithm

The closed phase WRLS-VFF-VT algorithm for speech analysis extracts the vocal tract parameters only from the glottal closed interval. The selection of the AR or ARMA model depends on the results of the V/U/M/N/S classification discussed in Chapter 4. This algorithm can be implemented as follows:

(1) Initialize the values of P_1, $\hat{\theta}_1$, λ_{min}, and E_1 and specify the filter order. (Experience shows that the values of P_1 and E_1 are insensitive to the algorithm provided they are adequately large, (e.g., $P_1 = 10^2$, $E_1 = 10^6$)).

(2) Compute the filter gain K_k, error covariance P_k and prediction error e_k using the WRLS-VFF algorithm as Table 3−1.

(3) Compute λ_k using eq. (3 − 32). If $\lambda_k < \lambda_{min}$, then $\lambda_k = \lambda_{min}$, where λ_{min} is given by eq. (3-34).

(4) Calculate the new filter coefficient vector $\hat{\theta}_k$ as per Table 3−1.

(5) Check for glottal closed phase using VFF based, LP error based , or DEGG based methods discussed in Chapter 4. If there is a closed glottal interval, then

extract the formants and their bandwidths by solving for the roots of the polynomial obtained from $\hat{\theta}_k$.

(6) Go to (2) until end of data.

5.2.3 Performance Evaluation of WRLS-VFF-VT Algorithm

The WRLS-VFF-VT algorithm is implemented and compared to block data processing algorithms: two-channel Closed Phase Covariance (CPC) (Krishnamurthy and Childers, 1985), modified Burg (mburg) (Kay, 1987), the modified Yule-Walker equations (mywe) (Kay, 1987), and the least squares modified Yule-Walker equations (lsmywe) (Kay, 1987).

Synthetic speech is used for our initial performance evaluation because we can specify and control such parameters as the formants, their bandwidths, and the excitation source with a formant synthesizer (Klatt, 1980; Pinto et al., 1989).

The isolated words and sentences spoken by a male subject is used to illustrate the performance of different algorithm such as two-channel CPC (Krishnamurthy and Childers, 1985).

The performance of each algorithm is evaluated according to its formant/antiformant tracking ability and its formant bandwidth estimation. For formant/bandwidth extraction, the roots of the numerator and denominator polynomials of the ARMA model are determined for the glottal closed phase interval for each period. All roots with a Q factor (center frequency divided by bandwidth) less than one were eliminated. The remaining roots are retained as potential formant roots. No other form of formant trajectory smoothing or filtering is done.

The analysis conditions for each method are indicated below.

1) For the AR model a 12th order was used.

2) For the ARMA model p=12 and q=6.

3) No preemphasis was used, since the closed phase vocal tract filters derived with and without preemphasis were virtually the same.

5.2.3.1 The generation of synthetic speech signal

We illustrate our results here with several synthetic speech signals generated by using a formant synthesizer (Klatt, 1980; Pinto et al., 1989) for the isolated nasal/nonnasal words and the sentence.

For synthesizing sustained phonations, such as sustained vowel /a/ and nasal sounds /m/, /n/, the amplitude, fundamental frequencies, and first four formant frequencies and bandwidths were used as constants as shown in Table 5−1. The frame size was also kept constant by 100 points (10ms interval).

For the synthesizing of the all-voiced multiple sounds utterance "We were away a year ago", the values of the above parameters are variable. The frame size is equal to the pitch-period. The formants frequencies and bandwidths are estimated from pitch synchronous LPC analysis and were smoothed by hand with WAVES+ DSP package.

Table 5−1 Formant(F)/antiformant(FZ) and bandwidth(B) used for generating synthetic speech signals (all frequencies are in Hz).

signal	F1	F2	F3	F4	B1	B2	B3	B4	FZ1	BZ1
/m/	390	1250	2150	3150	60	150	200	300	780	80
/n/	390	1250	2650	3950	60	150	200	300	1780	600
/a/	508	1069	2626	3035	102	48	131	213		

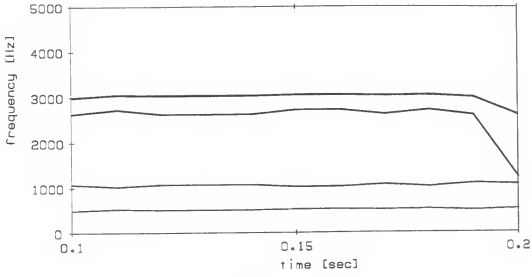
5.2.3.2 Experimental results

The estimated formant frequencies and bandwidths for synthetic speech /a/ using the WRLS-VFF-VT and the two-channel CPC are shown in Figure 5−1. The analysis results from the two-channel CPC [Figure 5−1(a) (c)] show that the bandwidths of formants cannot be accurately estimated due to the influence of input pulse, but the formant frequencies can be estimated accurately. Compared with the pitch-synchronous CPC method, the WRLS-VFF-VT can estimate the formant frequencies and bandwidths of the reference model more accurately [Figure 5−1(b) (d)].

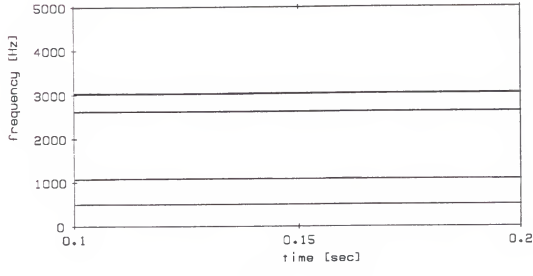
The spectra using FFT for the synthetic waveform /m/ are shown in Figure 5−2 (a). Figure 5−2 (b) and (c) show the formant and the antiformant tracks using the WRLS-VFF-VT algorithm. Figure 5−3 show the results of formant/antiformant tracks for the synthetic waveform /n/ using our algorithm. We see that the exact formants and antiformant based on the Table 5−1 can be estimated using the WRLS-VFF-VT algorithm.

Figure 5−4 shows the WRLS-VFF-VT algorithm experimentally compared with several spectral estimation methods which are modified Burg, modified YWE, and LSMYWE. In this figure we can see comparisons between the spectrum estimated by the proposed method and the spectra from other methods. We see that the spectra from other methods can approximate only the envelope of the FFT spectrum in Figure 5−4(a), and that the exact formants and the exact antiformant cannot be accurately estimated due to the influence of input pulse. As compared with other methods, the WRLS-VFF-VT can estimate ARMA spectra of the reference model accurately.

The analysis results for synthetic utterance, "We were away a year ago", are illustrated in Figure 5−5. A time varying reference model which has four formants as

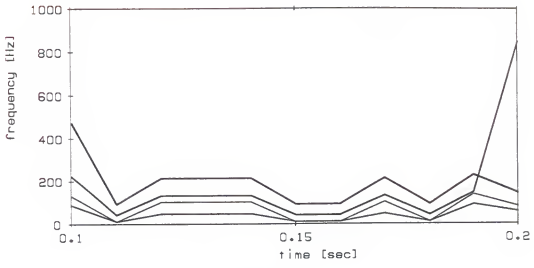


(a)

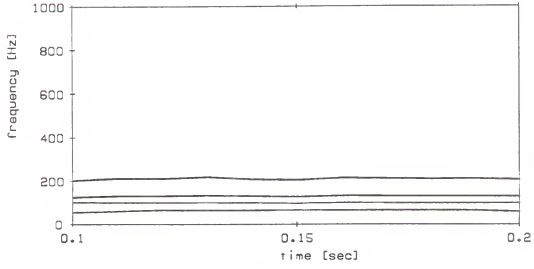


(b)

Figure 5−1. Formant frequency and bandwidth tracks for a synthetic speech, "a" (1st Freq.: 508.7 Hz, 2nd Freq.: 1069.0 Hz, 3rd Freq.: 2626.0 Hz 4th Freq.: 3035.0 Hz 1st B.W.: 102.7 Hz, 2nd B.W.: 47.95 Hz, 3rd B.W.: 131.8 Hz 4th B.W.: 213 Hz) using
(a)(c) 2-Ch CPC method, and (b)(d) WRLS-VFF method.



(c)



(d)

Figure 5−1. continued

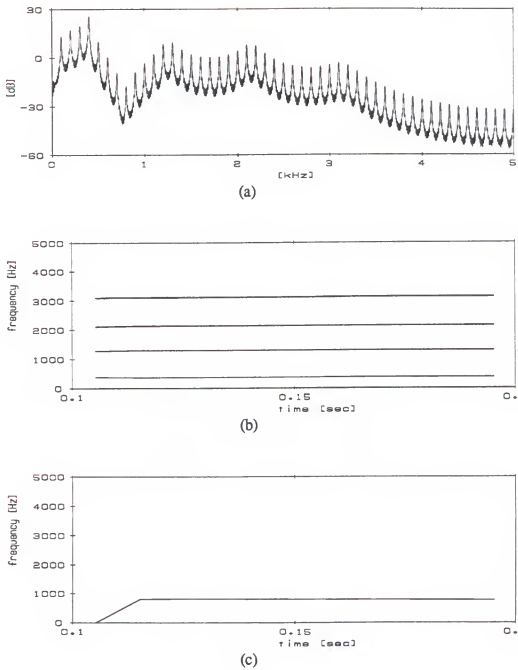
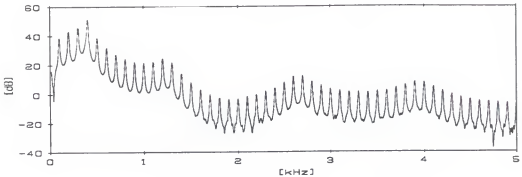
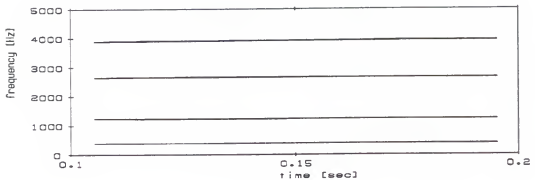


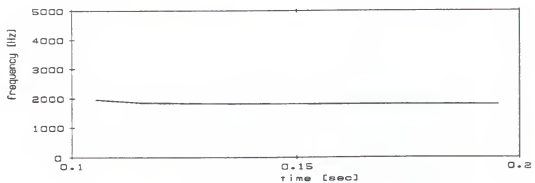
Figure 5–2. Formant/anti-formant tracking for the synthetic speech "m": (a) Spectrum, (b) formant tracking, and (c) anti-formant tracking.



(a)

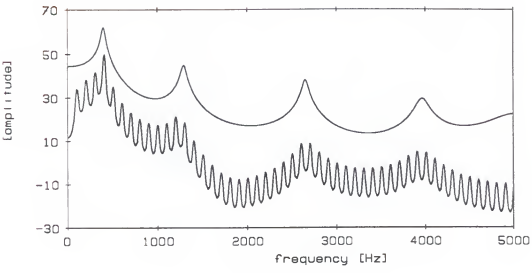


(b)

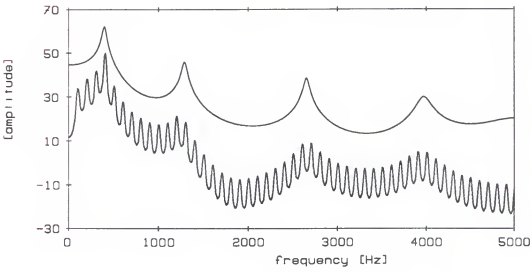


(c)

Figure 5–3. Formant/anti-formant tracking for synthetid speech "n": (a) spectrum, (b) formant tracking, and (c) anti-formant tracking.

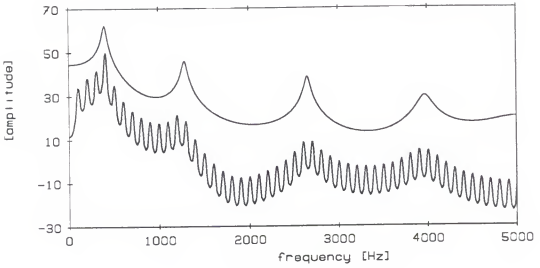


(a)

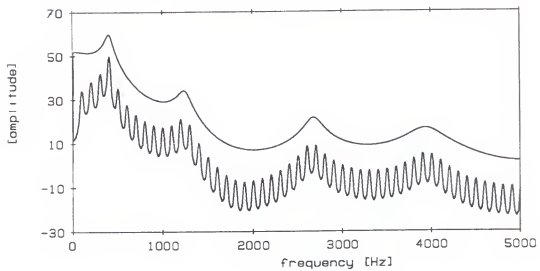


(b)

Figure 5−4. Comparison of spectral estimation for synthetic speech "n": (a) mburg, (b) mywe, (c) lsmywe, and (d) WRLS-VFF.

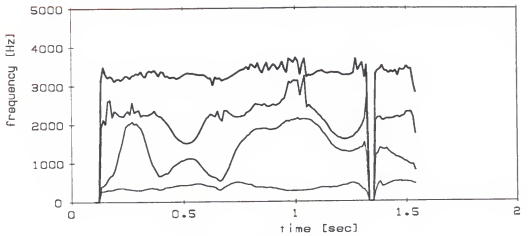


(c)

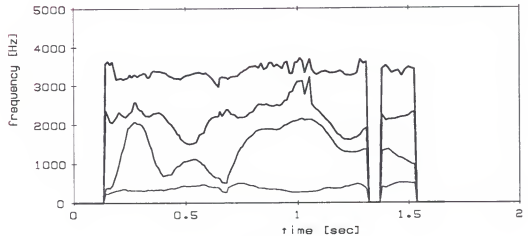


(d)

Figure 5–4 Continued



(a)



(b)

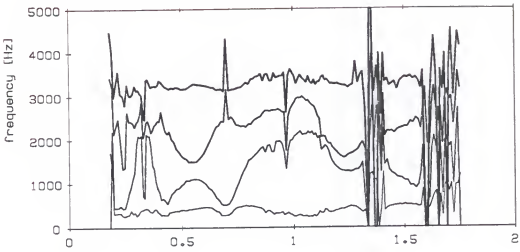
Figure 5−5. Formant tracking using synthetic speech for "We were away a year ago": (a) original formant (b) estimated formant by algorithm.

shown in Figure 5−5 (a) is used for the formant synthesizer. Figure 5−5 (b) illustrates the analysis result of the synthetic speech signal of this reference model. We see that formants are estimated accurately.

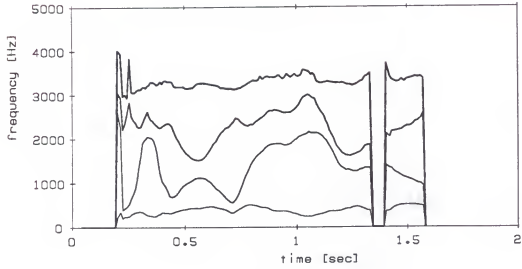
Figure 5−6 shows the formant tracking contours for the real speech of the utterance "We were away a year ago", estimated by the two-channel CPC and WRLS-VFF-VT methods, respectively. These data are for a male voice. An examination of the formant contours of this figure shows that the WRLS-VFF-VT method is better than the two-channel CPC in terms of tracking ability and smoothness of the tracks of formant estimates. The two-channel CPC method introduces abrupt discontinuities in the formant frequencies. The results show that the WRLS-VFF-VT method give the better results for the formant trajectories, in the sense that there are very few sudden jumps in the contours.

As the last example of the formant/antiformant tracking, we used the real speech signal of the nasalized utterance "Early one morning a man and a woman ambled along a one mile lane", spoken by a male speaker. Figure 5−7 shows the formant/antiformant tracking contours estimated by the WRLS-VFF-VT algorithm. A spectrogram can be used as the reference model for the formant/antiformant. The proposed algorithm provides smooth formant and antiformant tracks as shown in the figure. Furthermore, based on the spectrogram, both the formant and antiformant was stably and accurately estimated.

From the above test results we found that the block data processing techniques, such as the two-channel CPC, MYWE, and LSMYWE methods, gave reasonable estimates of the formants/antiformants and their bandwidths. However, the data windows used by these methods included the effects of the periodic excitation pulses, which affected the accuracy of the estimated formants and their bandwidths. However, the WRLS-VFF-VT method, which eliminated the influence of the pulse



(a)



(b)

Figure 5−6. Formant tracking for the real speech of "We were away a year ago" using (a) two-channe CPC (b) WRLS-VFF-VT.

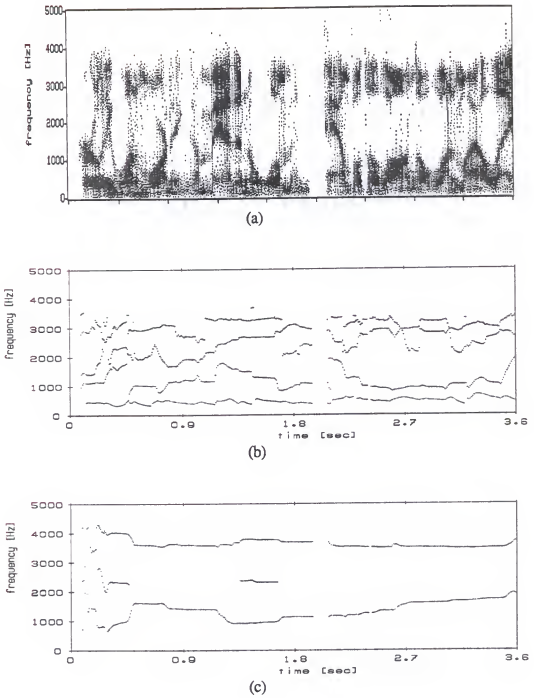


Figure 5–7. Formant and antiformant tracking for utterance "Early one morning a man and a woman amble the along a one mile lane": (a) spectrogram, (b) formant tracking, and (c) antiformant tracking.

excitation by using an input estimation as part of the algorithm, gave very accurate formant/bandwidth estimates and good spectral matching.

5.2.4 Summary

The algorithm has been tested extensively on considerable speech data from numerous subjects, including the data used in Chapter 4. Our results indicate that the closed phase WRLS-VFF-VT method of analysis seems superior to other methods. The method can be used for all-pole model analysis (e.g., vowels and diphthongs) as well as for pole-zero analysis (e.g., fricatives and nasals). The CPC method, which some consider superior to other LP methods, does not perform well when the closed phase interval is short, as with female or children's voices, or when the vocal tract characteristics change rapidly, as with the vowel/consonant transitions and some glide sounds. This is attributed to poorly estimated vocal track filter parameters obtained from a covariance matrix using a short data interval. Furthermore, matrix ill-conditioning problems may occur when solving the least squares equation in the CPC method.

The adaptive recursive algorithms derived from a least square cost function are known to converge rapidly (for short data records) (Haykin, 1985), and have an excellent capability to "track" an unknown parameter vector. In the proposed WRLS-VFF algorithm, a variable forgetting factor is used to allow the estimation process to track the time-varying parameters even more quickly.

5.3 Glottal Inverse Filtering

Voiced speech can be regarded as a deterministic glottal wave being input to a linear system, characterized by the vocal tract resonances. An interesting problem

is to analyze the speech waveform in order to separate it into vocal tract and glottal wave components.

Glottal inverse filtering is the process of making this separation. The speech signal is analyzed to determine the acoustic output response of the vocal tract to a glottal flow input. This response is used to specify the coefficients of an inverse filter which, applied to the acoustic speech wave, gives an estimate of the glottal flow wave as its output. The inverse filter serves to cancel the effect of vocal tract resonances from the speech signal to reveal the underlying voice source.

Glottal inverse filtering is important because the ability to determine the glottal flow wave has a number of major applications. One application is the improvement in quality of vocoded speech as used in speech synthesis and coded speech transmission. It has been established that source waveform pulse shape used to drive a vocoder has a significant influence on speech quality, and the use of pulse shape derived from theoretical analysis of the glottal flow wave results in improvement in synthesis quality over simple impulses.

Another application of inverse filtering is in noninvasive diagnosis of voice disorders. It may be possible to extract features of the inverse filter glottal wave estimate which indicate different types of abnormalities with the vocal folds. The limited amount of inverse filter data on abnormal voice indicates that the flow waveforms are difficult to interpret in the disordered case. The difficulty may stem from analysis artifacts, a problem which can be alleviated by improved inverse filter analysis algorithms. More recent work, however, indicates that although the flow waveform in the disordered case may be difficult to interpret, the emergence of a more normal inverse filter waveform can be observed in the course of voice therapy.

The type of information that one seeks to extract about the glottal source through the inverse filtering is the glottal pulse shape together with the cycle-to-cycle variation in the pulse shape. A parametric description of the pulse shape (e.g. duty

cycle, relative slopes of glottal opening and closing, pulse asymmetry) could form a parsimonious code for the voice source in a vocoder. The description could also form a set of features from which to make a voice disorder diagnosis.

5.3.1 Background for the Glottal Inverse Filtering

There are two approaches to making determination of the correct inverse filter: manual adjustment of filter coefficients (Rothenberg, 1973, 1977, 1981; Holms, 1976; Miller, 1959) and automatic determination of the inverse filter by way of LPC analysis (Wong et al., 1979; Berouti et al., 1977). This paper is focused on algorithms for conducting automated inverse filter analysis. Our several techniques of glottal inverse filtering, which are used to estimate the glottal flow waveform from the speech signal, are described below. Some examples of glottal inverse filtering as well as characteristics of the glottal flow waveforms will be discussed.

5.3.2 Glottal Wave Estimation

In order to study the acoustic characteristics of the glottal source we need to estimate or extract the glottal flow waveform from the speech signal. The process of estimating the glottal volume velocity by removing the vocal tract resonances from the speech signal is known as glottal inverse filtering. The objective of glottal inverse filtering is to determine the shape of the vocal fold waveform. Inverse filtering gives the true glottal flow provided that the filter is an exact inverse of the transfer function from the glottal flow to the speech wave. This implies that the frequencies and the bandwidths of the inverse filter should be set to represent glottal closed conditions. An accurate estimation of either the glottal volume velocity or the vocal-tract filter allows a determination of the other quantity, to within the limits of the assumed model.

5.3.3 Glottal Inverse Filtering

According to the linear source-filter theory of speech production (Fant, 1960), speech can be regarded as the result of the convolution of the glottal source waveform and the vocal tract transfer function. Figure 5−8 shows a block diagram representation of a linear speech production model. The glottal waveform is denoted by $u_g(n)$ and the output speech pressure waveform by s(n). For voiced speech the driving function, u(n), to the glottal shaping model, G(z), is a train of scaled unit samples. For unvoiced speech, the gain-adjusted, white gaussian noise is fed to the vocal tract filter directly, i.e., G(z) = 1. The vocal tract model, V(z), is assumed to be an all-pole model (Atal and Hanauer, 1971). Thus, the model is only an approximation for nasal sounds, which contains both poles and zeros. The speech pressure wave is related to the oral volume velocity at the lips through a radiation impedance R(z), that can be well represented by a simple zero (a high-pass filter) (Flanagan, 1972). Moreover, the lip radiation impedance effectively remains the same for different speech sounds.

For voiced speech it is possible to remove the effect of the glottal shaping filter by a glottal inverse filtering technique, if the instant of glottal opening and closure can be located accurately. From Figure 5−8, voiced speech, S(z), is represented as:

$$S(z) = U(z)G(z)V(z)R(z) = U_g(z)V(z)R(z) \qquad (5-1)$$

where

$$U_g(z) = U(z)G(z) \qquad (5-2)$$

Thus glottal inverse filtering can be conceptually defined as solving for glottal volume velocity $U_g(z)$, as can be seen in Figure 5−9 and Figure 5−10, by the equation

$$U_g(z) = \frac{S(z)}{V(z)R(z)} \qquad (5-3)$$

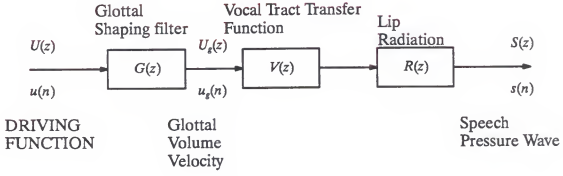


Figure 5–8. Block diagram representation of the linear speech production model.

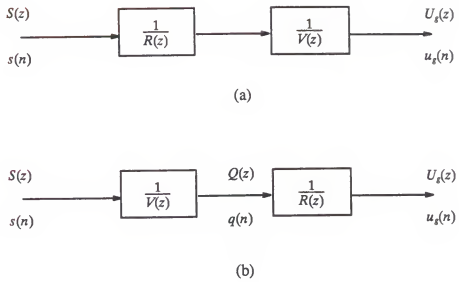


Figure 5–9. (a) Block diagram representation of the conceptualized glottal inverse filtering model, (b) Model obtained from (a) by interchanging the inverse vocal tract and the inverse lip radiation models.

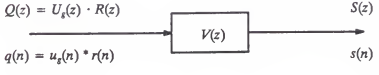


Figure 5–10. Equivalent representation of the linear speech production model shown in Figure 5–8 in terms of an effective driving function and the vocal tract transfer function. The symbol * denotes the convolution operation.

Figure 5−9-(a) shows the relationship between the glottal volume velocity, $u_g(n)$, and the speech pressure wave, s(n).

Since the lip radiation impedance is assumed to be the same for different speech sounds, the basic problem in the estimation of the glottal volume velocity waveform, $u_g(n)$, is to determine the parameters of the inverse filter, 1/V(z). Since the speech production model is linear, the lip radiation and vocal tract filters can be interchanged, leading to the arrangement of Figure 5−9-(b). By combining the lip radiation with the glottal excitation, an effective driving function, q(n), can be defined in the form

$$q(n) = u_g(n) * r(n) \tag{5−4}$$

or

$$Q(z) = U_g(z) \cdot R(z) \tag{5−5}$$

where the symbol * denotes the convolution operation. The effective driving function q(n) is equivalently the differentiated glottal volume velocity, since R(z) can be modeled by a simple zero.

Thus the linear speech production model depicted in Figure 5−8 can be equivalently described by the model in Figure 5−10. In terms of the effective driving function, q(n), the speech production model can be represented as

$$s(n) = -\sum_{i=1}^{M} a_i \, s(n-i) + q(n) \tag{5−6}$$

where a_i, i = 1, ..., M, is the coefficients of the vocal tract filter, V(z), modeled by an all-pole filter. During the interval of glottal closure the glottal volume velocity $u_g(n)$ is zero, and so is the driving function, q(n). Hence equation (5−6) becomes

$$s(n) = -\sum_{i=1}^{M} a_i \; s(n-i) \qquad (5-7)$$

Note that only one sample after the glottal closure instant (represented as nc), the speech waveform is strictly a function of the vocal tract resonances specified by a_1, ..., a_M, and the initial conditions s(nc), ..., s(nc-M+1). This result holds true over the entire closed glottal interval. The vocal tract filter parameters a_1,...,a_M, can be estimated by the covariance method of linear prediction (LP) analysis (Markel and Gray, 1976). Before applying the closed-phase covariance LP analysis, the closed interval must be identified.

5.3.3.1 GIF using WRLS-VFF-VT

From Figure 4−1, the positive peaks in the LP error function and the negative peaks in the VFF signal occur nearly simultaneously with the negative peaks of the differentiated EGG (DEGG) signal, which correspond to the instants of glottal closure (Childers et al., 1983; Childers and Krishnamurthy, 1985; Childers et al., 1990). From these observations, using these main pulses as indicators of glottal closure, a "pseudo closed phase" is selected as the analysis interval for pitch synchronous closed phase WRLS-VFF-VF analysis. This analysis estimates the vocal tract filter, which in turn is used to obtain the desired glottal volume-velocity waveform.

The basic ideas of detecting the closed phase intervals using the LP error based and the EGG based methods are similar to those of the two-pass CPC (Lee and Childers, 1989) and two-channel CPC (Childers and Krishnamurthy, 1985) methods. However, the first two methods use the variable threshold for the peak detection. Basically, we use the VFF based method described in Chapter 4, which does not require an auxiliary EGG signal as in the case of the two-channel method nor does it need a

second pass as in the case of the two-pass method. However, the LP error based or the EGG based methods can be used to detect the closed phase instants depending on the characteristics of input data.

To estimate the actual volume velocity, the filter coefficient for a single period is chosen as the adaptative process converges. The minimal data length for convergence of the WRLS algorithm is equal to twice the filter order (Haykin, 1985). The time of convergence can also be determined by checking the estimation error or the variable forgetting factor.

Having obtained the filter coefficients, minor adjustments are necessary to ensure that the inverse filter will only remove the formant poles from the speech signal. This step is needed because the estimated vocal tract model may have real poles at either zero frequency or the half-sampling frequency. Formants for the vocal tract, however, are always defined from only complex pole pairs. The real pole at zero frequency will typically occur due to low-frequency recording noise or a nonzero mean in the short analysis window. Real poles may also occur when the required filter order is over specified. The first effect is avoided by high-pass filtering the speech data, but the second effect may still lead to a real zero. If it is not removed, "jags" at the points of glottal closure will occur. A real pole may also occur at half the sampling frequency. When this pole has a narrow bandwidth, it generally indicates a formant location nearby, and thus should be retained. If a real pole occurs due to spectral shaping requirements in the analysis and there is not a nearby resonance, it will generally be of wide bandwidth. Including such a pole in the inverse filter will have a minimal effect on the results. Therefore, as a practical matter, any poles at half the sampling frequency are not removed. After eliminating any real roots near f=0Hz, a polynomial is reconstructed as the final inverse vocal tract model estimate.

The formant resonances of the vocal tract were estimated by solving the roots of the LP polynomial, and then shaping the formant structure by empirical rules, which

included: 1) discarding the roots with center frequencies under 250 Hz, 2) discarding the roots with a Q less than one, and 3) merging two adjacent roots. The refined formant resonances were then used to construct the vocal tract transfer function, which was used in the final GIF procedure. The direct output of the GIF operation is the differential glottal volume-velocity waveform. In order to remove any low-frequency trend from the glottal flow, any dc level of the differentiated glottal flow within the frame is removed. The resulting differentiated glottal flow is then normalized to have a unit power within the frame. A glottal volume-velocity is derived by carrying out an integration on the differential glottal v-v to cancel out the effect of the lip radiation.

A block diagram of the WRLS-VFF-VT method for GIF is shown in Figure 5−11.

Glottal source models that can be used to represent the glottal flow waveforms will be reviewed in the next section . Then a procedure for estimating parameters of a source model will be discussed. Applications of the analysis results to synthesizing and to automatically classifying the various voice types will also be discussed in the next chapter.

5.3.4 Voice Source Models

Based on the linear speech production model, source models have been used to represent the glottal flow pulse waveforms. A source model also can be used to describe the statistics of the glottal source flow characteristics. The glottal flow waveform has two components: (1) a residue component, and (2) a "noise" component. The residue component represents the main pulse of the glottal flow and can be described by a smooth function with only a few parameters. The noise component represents the remainder of the glottal flow. It includes the turbulent noise and the ripple component. While turbulent noise is mainly generated at a constriction in the

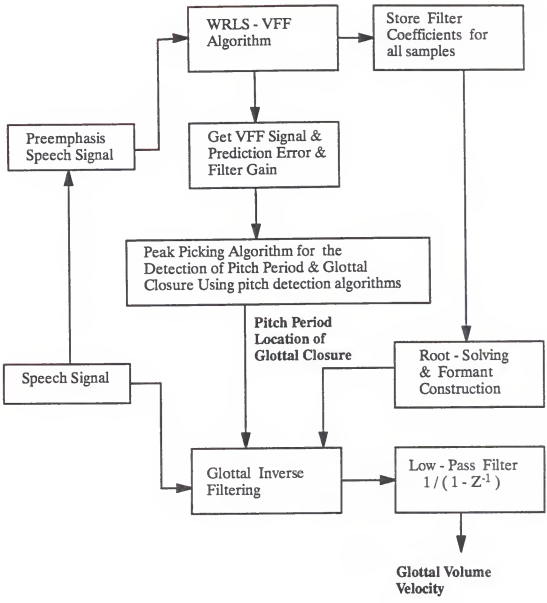


Figure 5−11. Diagram of the modified Ting method for glottal inverse filtering.

vocal tract, the energy stored in the vocal tract contributes to the ripple component. A model of the source should be flexible so that it may represent various types of glottal flow waveforms. A comparative study of some source models has been given by Fujisaki and Ljungqvist (1986). Here two typical glottal waveform models are discussed: Fant's model (Fant, 1979) and the LF model (Fant et al., 1985).

5.3.4.1 Fant's model

The wave shape of Fant's model (Fant, 1979) consists of rising and falling segments, as shown in Figure 5−12, and is represented by two equations:

$$U_g(t) = \frac{1}{2}U_o\left[1 - \cos\omega_g t\right] \qquad 0 \le t \le t_p \qquad (5-8)$$

$$U_g(t) = U_o\left[K\cos\omega_g(t - t_p) - K + 1\right] \qquad t_p \le t \le t_c \le T_o \qquad (5-9)$$

where t_c is the termination of the waveform. Given the fundamental frequency $F_o = 1/T_o$, where T_o is a pitch period, three basic parameters of Fant's model are the peak flow U_o, the glottal frequency $F_g = \omega_g/2\pi$, and the asymmetry factor (steepness factor) K. The rising segment reaches a peak value of U_o at $t = t_p$ and the falling segment has a value of zero at $t = t_c$. By applying these conditions to equations (5−8) and (5−9), respectively, the relationships between the time parameters and the waveform parameters can be obtained as follows:

$$t_p = \frac{\pi}{\omega_g} = \frac{1}{2F_g} \qquad (5-10)$$

$$t_c - t_p = \frac{1}{\omega_g}\cos^{-1}\left(\frac{K-1}{K}\right) \qquad (5-11)$$

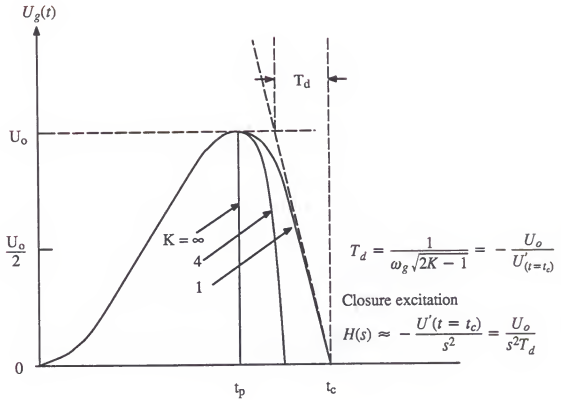


Figure 5–12. Fant's model of glottal wave.

or

$$K = \frac{1}{1 - \cos\omega_g(t_c - t_p)}$$

The termination of the waveform at t = t_c determines the primary excitation of the glottal wave. By differentiating equation (5−9) and applying the relationships of equation (5−11) and of $\sin(\cdot) = (1 - \cos^2(\cdot))^{1/2}$, under the condition of K > 0.5, the termination slope is given as

$$U'_{(t=t_c)} = \left[\frac{dU}{dt}\right]_{(t=t_c)} = -U_o \cdot \omega_g\sqrt{2K-1} = -\frac{U_o}{T_d} \qquad (5-12)$$

where

$$T_d = \frac{1}{\omega_g\sqrt{2K-1}} \qquad (5-13)$$

If K > 1 this gives the maximum slope during the course of the falling component. For 0.5 < K < 1 the maximum slope occurs prior to closure. At K = 0.5 the falling component is symmetrical to the rising branch. The lower bound of K is 0.5 in the present use of the model and represents a lowest degree of excitation strength. Generally the obvious shortcoming of a model with abrupt flow termination is that it does not allow for an incomplete closure or for a residual phase of progressing closure after the major discontinuity.

5.3.4.2 LF-model

Fant et al. (1985) proposed the LF model as shown in Figure 5−13. This model describes the differentiated glottal flow rather than the glottal flow itself. The differentiated flow is commonly used in speech synthesis, and includes the effect of radiation at the lips. The LF model consists of two segments. The first segment is an

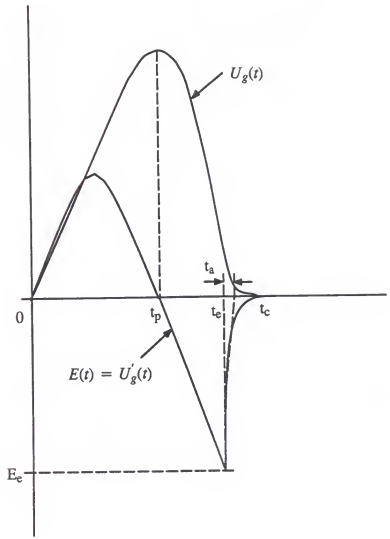


Figure 5–13. The LF-model of differentiated glottal flow $U_g'(t)$ - not drawn to scale.

exponentially growing sinusoid, and the second one is an exponential decaying function. Each segment is expressed as follows:

$$\frac{dU_g(t)}{dt} = E(t) = E_o \cdot e^{\alpha t} \sin \omega_g t \qquad 0 \le t \le t_e \qquad (5-14)$$

$$E(t) = -\frac{E_e}{\varepsilon t_a}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] \qquad t_e \le t \le t_c \le T_0 \qquad (5-15)$$

where T_0 is the pitch-period interval within which a waveshape of the LF model is defined. At time t_e both segments have the same value E_e. Besides the above relationships, there is a requirement of area balance which keeps the zero flow line from drifting. Thus the integral of the LF model time-function through the glottal period should vanish, i.e.,

$$\int_0^{T_o} E(t) = 0 \qquad (5-16)$$

The three parameters of the first segment of the LF-model are:

(1) E_o which is a scale factor.

(2) α = -Bπ where B is the "negative bandwidth" of the exponentially growing amplitude.

(3) $\omega_g = 2\pi F_g$ where $F_g = 1/2t_p$ and t_p is the rise-time (the time from glottal opening to maximum flow).

In the second part of the LF model, the parameter t_a is the time constant of the exponential curve and is determined by the projection on the time axis of the derivative at time t_e, at which the negative peak of the LF model occurs. The parameter E_e is the negative amplitude of the excitation spike at time t_e. The parameter t_c is the moment when complete closure is reached. The parameter ε is the decay constant of the recovery phase exponential. The basic four parameters E_o, α, ω_g, and ε are called the

"direct synthesis parameters" of the LF model, while the time parameters t_p, t_e, t_a, t_c are called the "timing parameters". Each LF model timing/direct-synthesis parameter can be thought of as independent of one another, because a unique combination of timing/direct-synthesis parameters can generate a unique waveshape.

The first segment of the LF model represents the differentiated flow from glottal opening to the instant when the main excitation occurs (the moment of maximum discontinuity in the glottal airflow function, which normally coincides with the moment of the maximum negative flow derivative). The second part of the model is an exponential segment that allows a residual flow (dynamic leakage), from the point of maximum closing discontinuity at time t_e towards maximum closure, when the vocal folds close at time t_c. The effect of the return phase on the source spectrum is, due to its exponential waveshape, approximately a first order low-pass filter with a cutoff frequency $F_a = 1/(2\pi t_a)$ (Fant and Lin, 1988). This means that the longer the return phase, the lower the cutoff frequency, and the greater the reduction of the high frequency energy.

The LF model time-function is generated by using the direct synthesis parameters, i.e., E_o, α, ω_g, and ε. However, for many research applications, such as model fitting to inverse filtered glottal flow waveforms, it is easier to specify the timing parameters - t_p, t_e, t_a, t_c - and E_e rather than the direct synthesis parameters. The direct synthesis parameters can be easily computed from the timing parameters and E_e. The procedure to obtain the corresponding direct synthesis parameters from the timing parameters and E_e is as follows:

(1) The intermediate parameter ε can be determined by an iterative procedure from equation (5−15) by letting $t = t_e$, i.e., from

$$\varepsilon t_a = 1 - e^{-\varepsilon(t_c - t_e)} \tag{5-17}$$

For small values of t_a, ε is approximately equal to $1/t_a$.

(2) By definition, $\omega_g = \pi/t_p$.

(3) The solution for the parameter α can be obtained by applying the area balance constraint of equation (5−16) with the solution of E_o from equation (5−14)

$$E_0 = -\frac{E_e}{e^{\alpha t_e}\sin\omega_g t_e} \tag{5-18}$$

In estimating the LF model parameters, the parameter t_c, which represents the closing instant, is usually set to T_o, the time of glottal opening for the following pulse period. This implies that the model may lack a closed phase. In practice, however, for small values of t_a the exponential function of the second part of the LF model will have negligible value and thus provides an effective closed phase.

The LF-model function is continuous until the main excitation, and therefore does not introduce additional excitation at the flow peak. In comparison, Fant's model consists of two different segments, a rising segment up to maximum flow and a falling segment down to complete closure. The discontinuity between the two segments introduces a secondary weak excitation at the flow peak. The major difference between these two models is that the LF model allows for a residual phase of progressive closure, while the Fant model always generates an abrupt closure. The existence of the residual closing phase in the LF model gives the flexibility of modeling various voice types more efficiently.

In summary, the LF model is a good approximation for non-interactive flow parameterization in the sense that it ensures an overall fit to commonly encountered

glottal flow wave shapes with a minimum number of parameters (Fant et al., 1985). It is flexible in its ability to match various phonations, for example, breathy voice. The four glottal factors important for characterizing several voice types (Lee and Childers, 1989) are: (1) glottal pulse width, (2) glottal pulse skewness, (3) abruptness of glottal closure, and (4) turbulent noise. The first three glottal factors can be modeled effectively with the LF model. The fourth factor should be incorporated separately to provide a complete model. In this research the LF model, with an accommodation for turbulent noise, was used to parameterize the waveform characteristics of the glottal flow.

5.3.5 Modeling of the Glottal Flow Waveform

For voiced sounds, the closed phase vocal tract parameters were calculated from the closed phase WRLS-VFF-VT method. Then the glottal flow waveform was obtained by inverse filtering. A source model was fitted to the glottal flow waveform to obtain the glottal source parameters. The glottal opening instant and the duration of the glottal open phase as well as the closing instant and the closed phase duration were determined from the differentiated EGG signals.

5.3.5.1 Measurement of model parameters

We used the LF model to extract the parametric features of the glottal flow waveform. When fitting the LF model to inverse filtered differentiated glottal flow waveforms, the parameter t_c, which represents the closing instant, is assumed to be equal to T_o, the time of glottal opening for the next pulse period. In this study, we defined the parameter t_c as the instant at which the modeled differentiated glottal flow amplitude drops to 1% of its peak value and was computed from the matched model waveform. Thus, the parameter t_c does not represent the actual closing instant of the

LF model; rather it can be considered as a settling time of the model. We used the parameter t_c as an approximate to the closing instant of the LF model.

Figure 5−14 shows the block diagram for the algorithm that matches the LF model to the measured differentiated inverse filtered waveform. Within an analysis frame (corresponding to one pitch period) of the inverse filtered differentiated glottal flow waveforms, the values of t_e and E_e parameters are easily measured. The parameter t_a is determined by a least square error (LSE) criterion between the inverse filtered differentiated glottal flow waveform and the closing part of the LF model given by equation (5−15). Then possible candidates for the parameter t_p are located in the range from the instant of the glottal opening to the instant t_e. For each candidate of t_p, the direct synthesis parameters - E_o, α, ω_g, and ε - are calculated. Based on the direct synthesis parameters obtained, the modeled differentiated glottal flow waveforms are generated and the total squared errors are computed. Finally the parameter set which gives the minimum total squared error is selected as the best matching LF model for the frame.

The timing parameters of the LF model are closely related to the glottal waveshape factors, for example, t_c to glottal pulse width, t_a to abruptness of glottal closure, and t_e to the instant of the main excitation during glottal open. To describe quantitatively the waveshape characteristics of the modeled glottal flow, we defined the open quotient (OQ) and the speed quotient (SQ) for the LF model waveshape. The open quotient (OQ_{LF}) of the LF model waveshape is defined as the ratio of the open phase to the pitch period, i.e.

$$OQ_{LF} = \frac{open\ \ phase}{pitch\ \ period} = \frac{t_c}{T_o} \qquad (5-19)$$

The range of values for the open quotient is from 0 (no open phase) to 1 (no closed

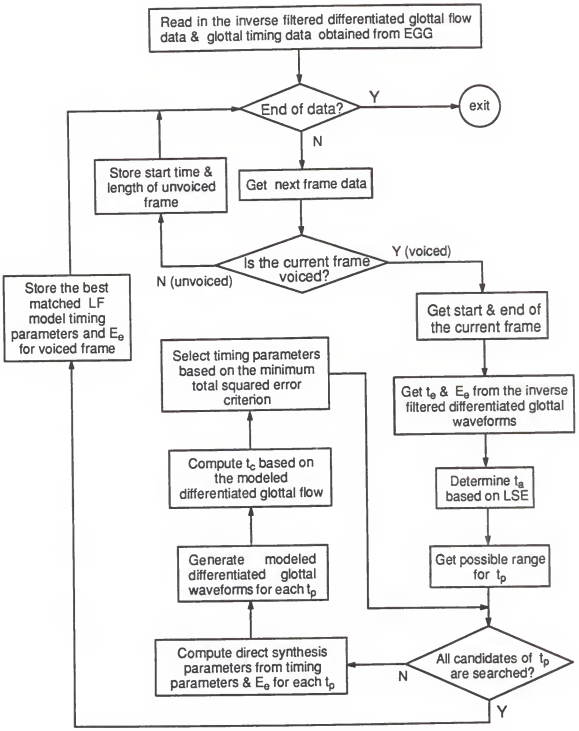


Figure 5−14. The block diagram for the LF model matching algorithm.

phase). The speed quotient (SQ_{LF}) of the LF model waveshape is defined as:

$$SQ_{LF} = \frac{opening\ \ phase}{closing\ \ phase} = \frac{t_p}{t_c - t_p} \qquad (5-20)$$

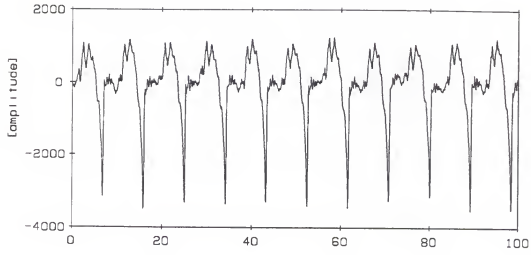
Glottal pulse skewness is commonly represented by the speed quotient. Values of the speed quotient can range from 0 (no opening phase) to infinity (no closing phase). Both extreme values of the speed quotient cannot occur due to the physiological limitation of human articulatory movements. Note that in equation (5−19) and (5−20), the computed t_c is used to approximate the closing instant. The measured data for both the OQ and the SQ from the matched LF model waveforms are reported in a later section.

5.3.6 Experimental Results of GIF and the LF Model Matching

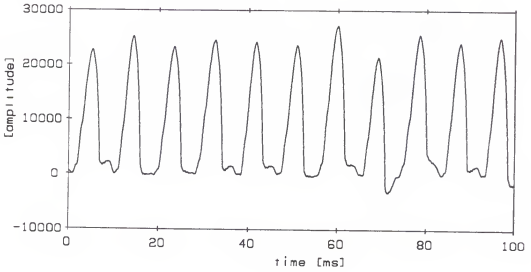
The glottal flow waveform and its differentiated waveform obtained from the glottal inverse filtering of the sustained vowel, /a/, from a male speaker (DMH) are shown in Figure 5−15. Here, we can see that the speaker has a good closed phase. The modeled differentiated glottal flow and the modeled glottal flow for the vowel /a/ (DMH) in Figure 5−15 are shown in Figure 5−16.

The resulting glottal v-v waveform estimate, obtained by integrating the differentiated glottal v-v, has a relatively slow rise time at the onset of the glottal opening and is seen with a small amount of superimposed ripple. This behavior for speech production with moderate intensity is typical and has been observed by most researchers in this area for moderate intensity vowel phonation (Miller, 1959; Lindqvist, 1970; Rothenberg, 1973; Wong et al., 1979).

To verify the glottal inverse filtering and the model matching, a synthesized speech sound with known source input was used to test the program. The input to the

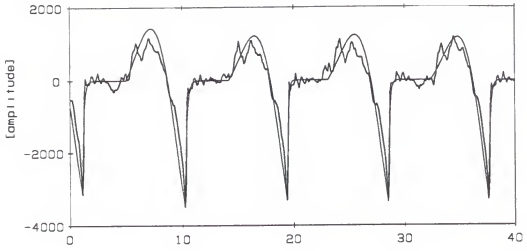


(a)

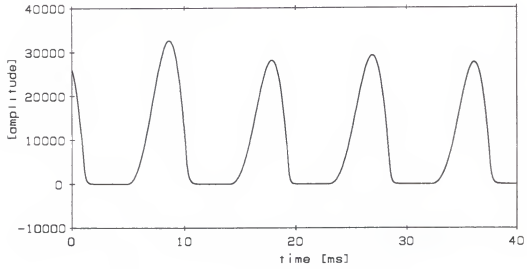


(b)

Figure 5–15. Glottal inverse filtering result for sustained vowel, /a/, from a male speaker (DMH): (a) differentiated glottal volume velocity and (b) glottal volume velocity.



(a)



(b)

Figure 5–16. LF-modeled data for the vowel /a/ in Figure 5–15: (a) normalized differentiated glottal flow and the modeled waveform, (b) glottal flow.

glottal inverse filtering was generated by concatenating two synthesized sustained vowels, /a/ , with 20 (ms) transient between them. To synthesize the sustained vowel, the input was a periodic sequence of LF modeled glottal pulses with a zero return phase. The pitch frequency of the glottal pulses was 100 Hz. The synthesized speech waveform for /a/ is shown in Figure 5−17 (a). Figure 5−17 (b) shows the VFF signal waveform which gives the information of glottal closed instant. The two waveforms, the differentiated glottal v-v and glottal v-v, obtained from the GIF algorithm are shown in Figure 5−17 (c) and (d). The modeled differentiated glottal flow and the modeled glottal flow waveforms by the model matching algorithm using the LF model parameters from GIF are shown in Figure 5−17 (e) and (f). Figure 5−17 (g) and (i) show the original differentiated glottal v-v and the original glottal v-v for the input source of synthesizer. We can be confident from the results that the glottal inverse filtering and the model matching algorithms work properly.

Examples of the inverse filtering and the LF modeling of sentences, "We were away a year ago" "Should we chase those cowboy" and "That zany van is azure" spoken by a normal male subject DMH, are shown in Figure 5−18, NO TAG, and NO TAG. Note that the inverse filtered glottal flow waveforms for these sentences contain unvoiced, mixed and silent segments. It is not easy to estimate accurately the opening instants for the modeled differentiated glottal flow from the differentiated EGG. Thus, a less than perfect match between the real and modeled differentiated glottal flow may result. A possible solution is to adjust the model opening instant until an optimum, in the sense of the minimum total squared error, is found. This procedure, however, requires a considerable amount of computing time.

5.3.7 Comparison of Different GIF Methods

The performance of two GIF methods, two-channel CPC and WRLS-VFF-VT, were evaluated using a sentence, "We were away a year ago", spoken by a normal male

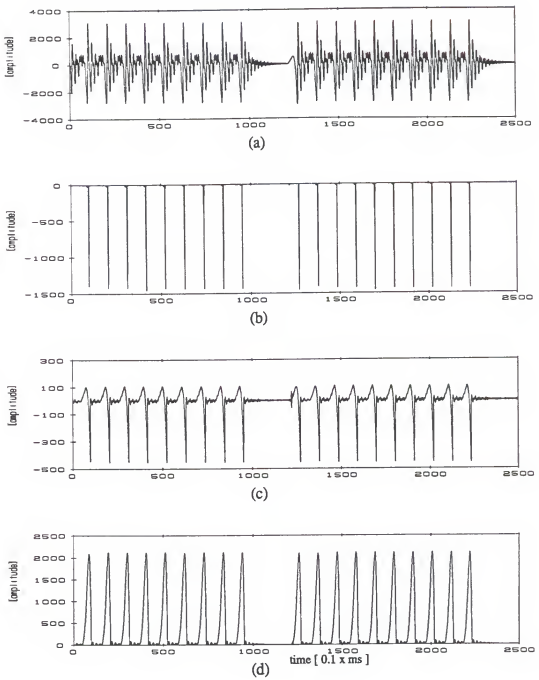
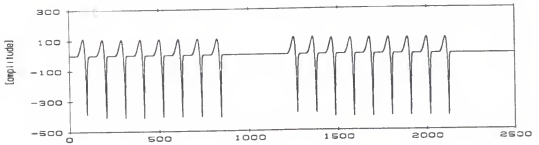
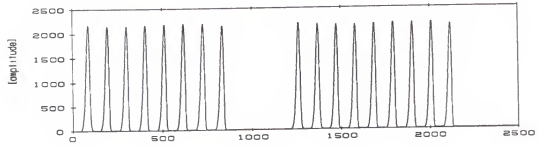


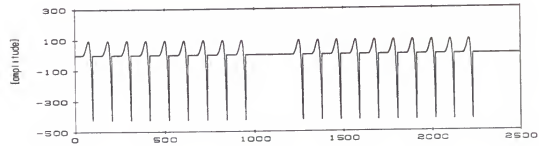
Figure 5−17. Analysis results for synthetic speech signal/a/: (a) synthetic speech, (b) VFF, (c) differentiated glottal v-v from GIF, (d) glottal v-v from GIF, (e) modeled differentiated glottal v-v by model matching, (f) modeled glottal v-v by model matching, (g) original differentiated glottal v-v, and (i) original glottal v-v.



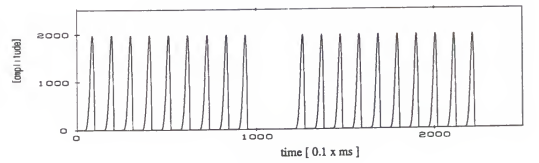
(e)



(f)



(g)



(h)

Figure 5–17. continued

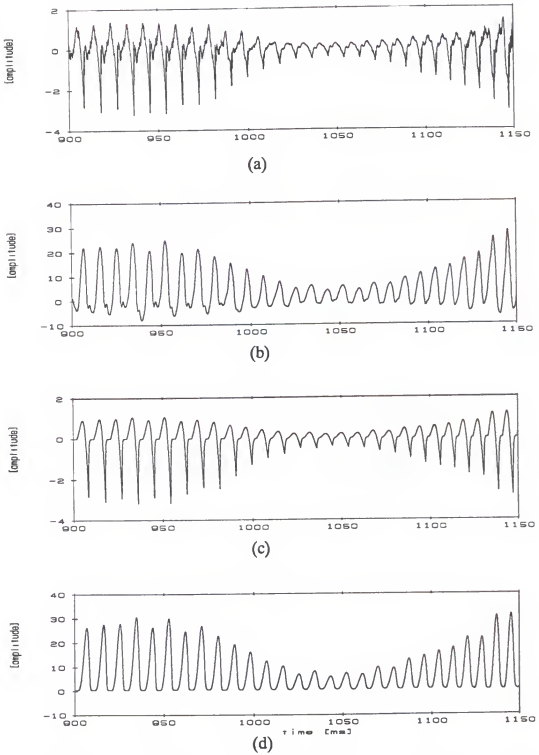


Figure 5−18. GIF for a sentence,"We were a way a year ago": (a) Glottal waveform, (b) Differentiated Glottal waveform, (c) LF model, and (d) LF model.

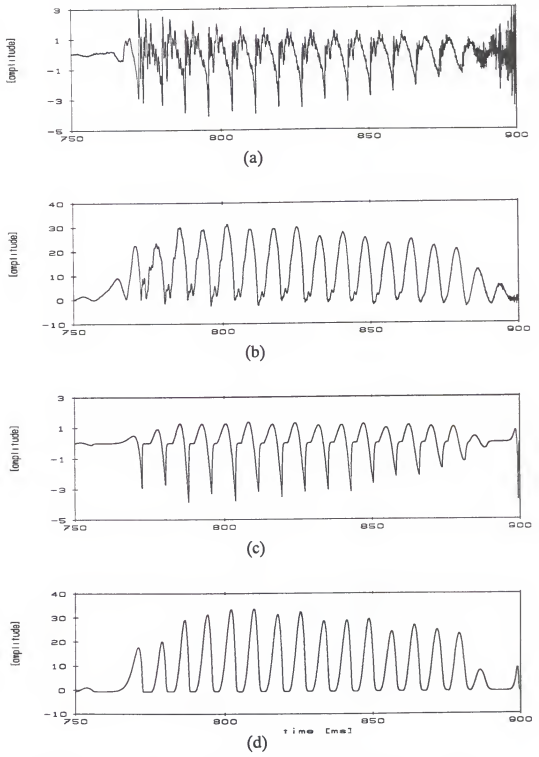


Figure 5−19. GIF for a sentence,"Should we chase those cowboy": (a) Glottal waveform, (b) Differentiated Glottal waveform, (c) LF model, and (d) LF model.

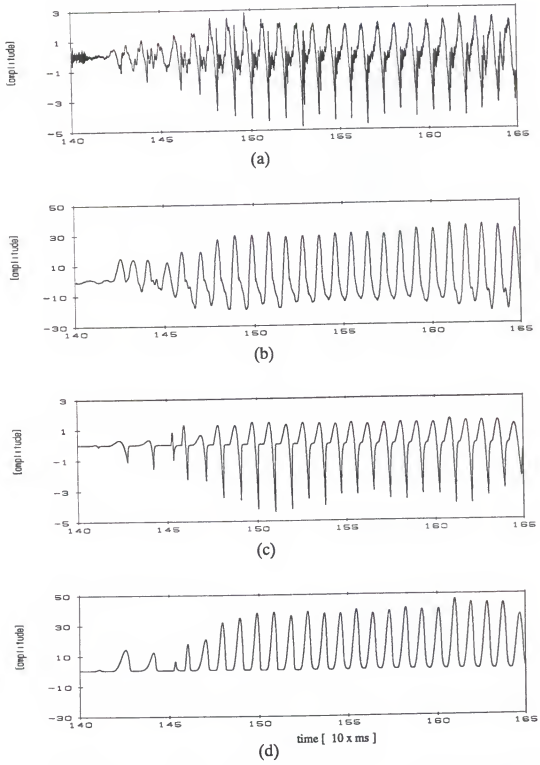


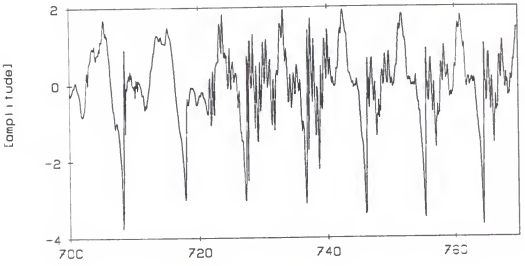
Figure 5–20. GIF for a sentence,"That zany van is azure": (a) Glottal waveform, (b) Differentiated Glottal waveform, (c) LF model, and (d) LF model.

speaker. Figure 5−21 (c)-(a) and Figure 5−21 (d)-(b) show the estimated glottal v-v waveform and its derivative for the two-channel CPC and the WRLS-VFF-VT methods, respectively. For the two-channel CPC method, some high frequency components appear in the closed phase region, presumably due to an incomplete cancellation of the pitch epoch from the vocal tract transfer function. For the WRLS-VFF-VT method, a continuous glottal v-v waveform was obtained. The waveform agreed with the expected characteristics for the glottal excitation source such as a flat closed region and a sharp slope at closure.

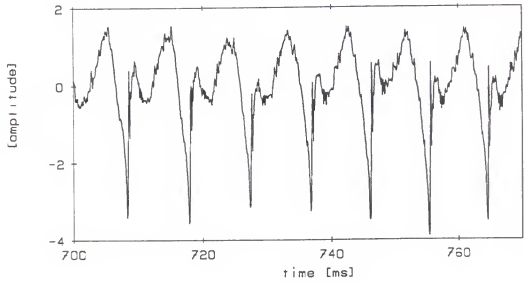
The two-channel CPC and WRLS-VFF-VT methods have also tested for the pathological voice, vocal fry. The result is shown in Figure 5−22. It is known that vocal fry has a very distinct closed phase compared with that of a normal voice. The WRLS-VFF-VT method can give a better estimate of the glottal v-v waveform than that of the two-channel CPC mehod.

5.3.8 Summary

A WRLS-VFF-VT method for the GIF has been presented in this section to estimate the glottal v-v waveform by the inverse filtering of the acoustic speech signal. It can provide reliable glottal v-v waveform estimates automatically for the normal and pathological speech signals. This method does not require an auxiliary EGG signal as in the two-channel CPC method. We compared our method to the two-channel CPC method using both normal and pathological speech signals and using both real and synthetic speech signals to evaluate the performance. A major benefit of our method is that it is able to provide interpretable results on higher fundamental frequency speech such as with females. Female speech is notoriously difficult to analyze using LPC methods as well as is the speech that doesn't have a glottal closed phase.

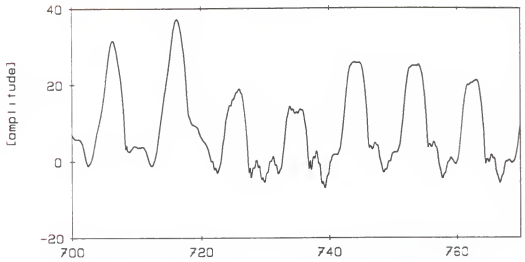


(a)

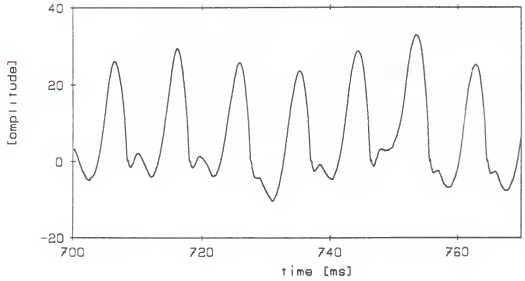


(b)

Figure 5−21. Comparison of glottal inverse filtering for a sentence "We were away a year ago": (a) & (c) Normalized differentiated glottal flow waveforms for CPC (b) & (d) glottal flow waveforms obtained by integrating (a) & (c), respectively for WRLS-VFF.



(c)



(d)

Figure 5−21. Continued.

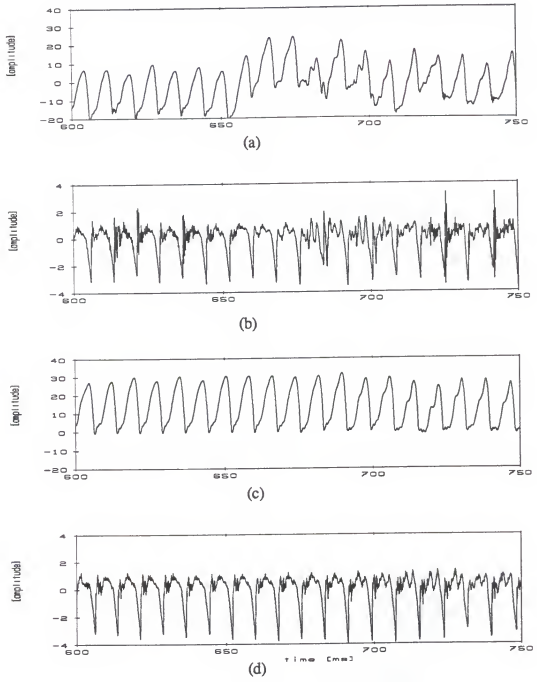


Figure 5−22. GIF for the pathological voice /i/ spoken by JTO using different methods: (a) glottal v-v (b) D-glottal v-v using the two-ch. CPC method, and (c) glottal v-v (d) D-glottal v-v using the WRLS-VFF-VT method.

CHAPTER 6
SYNTHESIS AND APPLICATION

In this chapter, we discuss synthesizing by the analysis results using a flexible formant synthesizer (Lalwani, 1991) to evaluate our analysis algorithms. Figure 6−1 shows the block diagram of the analysis procedure for synthesis. In addition, we discuss the several synthesis strategies that were employed for synthesizing various types of speech utterances using the flexible formant synthesizer. Application of the analysis results to the automatic classification of different voice types will also be discussed.

6.1 Synthesis Strategies and Experiment Results

The block diagram of the flexible formant synthesizer architecture (Lalwani, 1991) is shown in Figure 6−2. In the following section we outline some of the basic procedures used to synthesize the speech signals using the default configuration of the flexible formant synthesizer (Lalwani, 1991). One should realize that it is not possible for us, in this study, to outline the general synthesis strategies, such as those used for the synthesis of English syllables, words, etc., in the rule-based speech synthesis systems. The strategies outlined here are simple steps to be followed to synthesize simple utterances. The exact synthesizer parameters and synthesis procedures depend upon the speech tokens to be synthesized and the context in which such speech tokens will be used. For example, if high-quality synthetic speech is desired, then the synthesizer parameters should be accurately specified and the synthesis procedure should be carefully controlled. In the following sub-sections, we describe how the

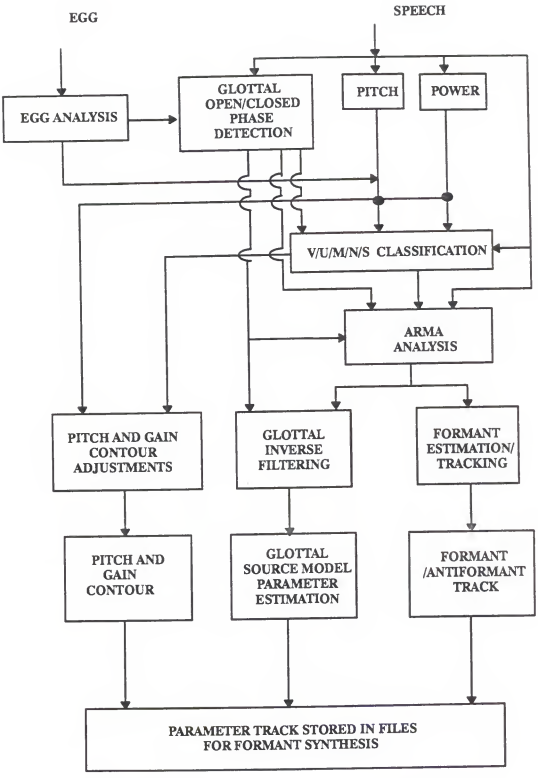
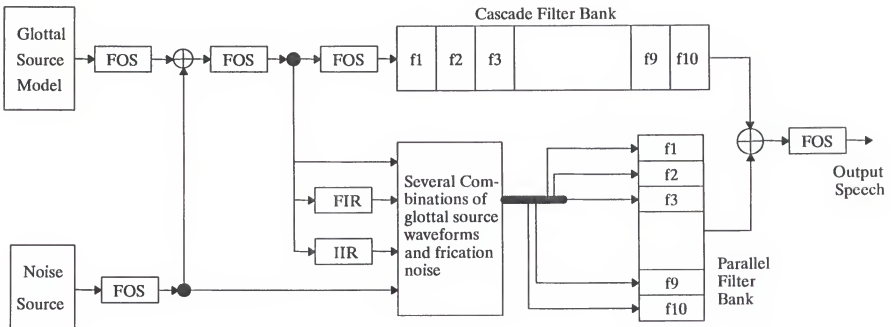


Figure 6–1. Block diagram of the analysis procedure for synthesis.



FOS : First Order System (which may be a FIR filter, an IIR filter or a by−pass path)
f1 ··· f10 : Second Order Systems (which may be a resonators, an an anti−resonators or multipliers)

Figure 6−2 Configuration of the flexible formant synthesizer

excitation source (glottal and noise) are generated and the filter banks (cascade and/or parallel) are configured to synthesize voiced, unvoiced and mixed sounds.

6.1.1 Voiced Sounds (Lalwani, 1991)

A glottal source model is used as an excitation source for synthesizing vowels (/i/, /ɪ/, /ɛ/, /æ/, /a/, /ɜ/, /ʌ/, /ɔ/, /u/, /ʊ/ and /o/), semi-vowels (/w/, /l/, /r/ and /j/), diphthongs (/aɪ/, /ɔɪ/, /aʊ/, /eɪ/, /oʊ/ and /juʊ/) and nasals (/m/, /n/ and /ŋ/). The glottal source model generates the glottal source pulses that simulate the volume-velocity pulses of air produced by the quasi-periodic vibrations of the vocal folds as the air flows from the lungs to the pharynx. The shape of the glottal pulses is controlled by the parameters of the glottal source model. The shape of the glottal pulses determine the vocal characteristics of synthesized speech, such as breathy, normal, etc. A detailed explanation of how to control the values of the parameters of the glottal source models (discussed in Appendix B) to obtain various glottal pulse shapes is beyond the scope of this section.

In the flexible formant synthesizer, the rate of generation of the glottal pulses (fundamental frequency) is specified by the parameter "f0." The higher the value of the parameter "f0," the higher is the value of perceived "pitch." The energy/power in each of the glottal pulses is specified by the parameter "av." The higher the value of the parameter "av," the higher is the perceived "stress." The duration of a voiced sound is determined by the number of consecutive frames for which both the "av" and "f0" parameters are specified to be nonzero. The gain and the pitch contours for an utterance represent the variation in the values of the "av" and "f0" parameters during the utterance. It is the pitch contour, gain contour and the duration of each sound in an utterance that determines the "intonation" and "stress" patterns of an utterance.

According to the acoustic theory of speech production, the non-nasalized voiced sounds can be represented by an "all-pole" vocal-tract transfer function. Different

voiced sounds are produced by different vocal tract configurations and movements of articulators, and therefore, have different formant parameters (frequencies, bandwidths and amplitudes). In the American English language, most voiced sounds can be adequately represented by the lower three formant frequencies and bandwidths. Over the duration of an utterance, there is a wide range of variation in the values of the formant frequencies and amplitudes but only slight variation in the values of formant bandwidths. Also, there is variation in the inherent duration of each voiced sound. Typical values of formant frequencies and bandwidths for the voiced sounds have been established by analysis of several speech tokens for males and females (Peterson and Barney, (1952); Klatt, 1980; Childers and Wu 1990). Also, a table of minimum and inherent durations of the voiced sounds is given in Klatt (1987). Several researchers have shown that a fairly good copy of an all voiced utterance can be obtained by replicating the formant tracks (formant frequencies and bandwidths) of the first three formants, and the pitch and gain contours.

For synthesizing sustained phonations, such as sustained vowels /a/, /ɪ/, etc., the "av," "f0," "f1," "f2," "f3" "b1," "b2" and "b3" parameters should be specified as constant. For synthesizing an all-voiced multiple sounds utterance the values of the above parameters are variable and are specified through parameter tracks. Default values of the fourth and fifth formant frequencies and bandwidths can be used during the synthesis. If the parallel filter bank is used to synthesize an all-voiced utterance, the amplitude control parameters of all the resonators in the parallel filter bank should be set to zero, since they are automatically determined (to make the magnitude frequency response of the parallel filter bank equivalent to that of the cascade filter bank), unless variations in the "normal" amplitude of the formants is desired. The frame size may be kept constant, or can be equal to the pitch-period, if the pitch synchronous synthesis flag is set to one. The total number of frames to be synthesized

is equal to the length of the parameter tracks. The values in the parameter tracks are used to update the values of these parameters at the beginning of each frame.

When the cascade filter bank is used to synthesize an utterance with the nasal sounds, a parameter track for the center frequency of the nasal anti-resonator should be specified. This parameter track should have a constant value of 250 Hz (default value) except for the duration of a nasal murmur or a nasalized vowel. For the duration of the nasal murmur, the center frequency of the nasal anti-resonator should be shifted to 450 Hz (Klatt, 1980). For the duration of nasalized vowel, a "pole-zero-pole" combination is created using nasal resonator, nasal anti-resonator and first formant generator (resonator). The first formant frequency should be increased by 100 Hz and the center frequency of the nasal anti-resonator should be equal to the average value of the the center frequency of the nasal resonator (fixed at 250 Hz) and the first formant frequency (Klatt, 1980). When using the parallel filter bank and Klatt's rules (with either Klatt's scale factors or the new scale factors) for synthesizing the nasal murmur and nasalized voiced sounds, the amplitude control parameter of the nasal resonator has to be set to about 60 dB for the duration of the nasal sound and zero dB otherwise. When using the new procedure the amplitude control parameter of the nasal resonator should be kept zero since the amplitude of the nasal formant is set equal to the amplitude of the nasal resonator in the cascade filter bank. For both the procedures, the amplitude control parameter of the other resonators should be kept zero, unless variations in the "normal" amplitude of the formants is desired.

In the cascade/parallel synthesizer configuration only the cascade filter bank is used for synthesis of a sustained vowel and an all voiced sentence. Therefore, we have created an additional synthesizer configuration, "all-cascade synthesizer configuration," in which only the cascade filter bank is used for synthesis. Figure 6−3(a) - (d) show the analysis results of the V/U/M/N/S classification, the gain contour, the fundamental frequency, and the formant tracking for the sentence "We

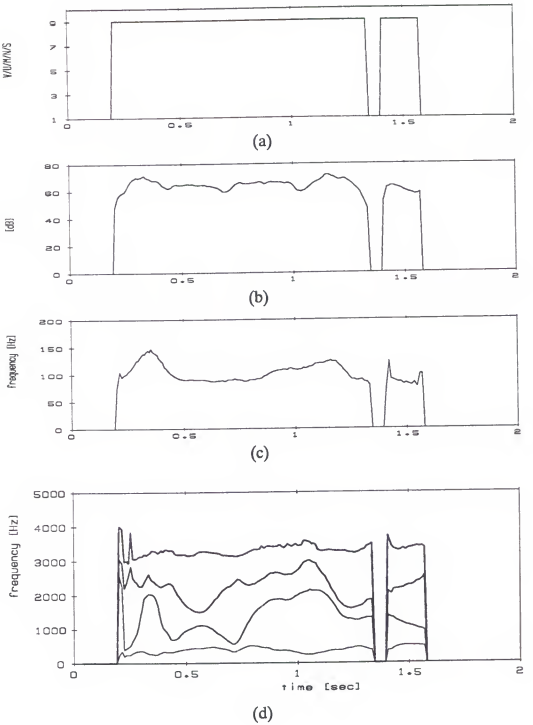


Figure 6–3. Parameter tracks for the sentence "We were away a year ago": (a) V/U/M/N/S, (b) gain contour, (c) fundamental frequency, and (d) formant frequency tracks

were away a year ago." The spectrograms of the natural and synthesized utterance of this sentence are shown in Figure 6–4 (a) and (b).

The cascade/parallel synthesizer configuration is used for synthesis of nasals and nasalized vowels. Figure 6–5 (a) - (e) and Figure 6–6 (a) - (e) show the analysis results of the V/U/M/N/S classification, the gain contour, the fundamental frequency, the formant tracking, and the antiformant tracking for the sentence "Early one morning a man and a woman ambled along a one mile lane." and for the sentence "That zany van is azure", respectively. The spectrograms of the natural and synthesized utterance of two sentences are shown in Figure 6–7 and Figure 6–8.

A visual comparison of the spectra for these two sentences show a good match between the synthesized and natural speech utterances.

6.1.2 Unvoiced Sounds (Lalwani, 1991)

The excitation source for synthesizing unvoiced sounds is a white-noise source. The white-noise source models the turbulent noise produced by a narrow constriction or an occlusion in the vocal tract during the production of unvoiced sounds. When the constriction in the vocal tract is narrow, the steady air flow from the lungs becomes turbulent, producing friction-like noise which is the source of fricatives (/f/, /θ/, etc.). When the air pressure that builds-up behind an occlusion in the vocal tract is suddenly released, there is a brief interval of frication (due to the sudden turbulence of escaping air) followed by a period of aspiration (steady air flow from the open glottis). The parameter "af" specifies the gain of the noise source during the synthesis of fricatives. The parameter "ah" specifies the gain of the noise source during the synthesis of an aspiration. The synthesizer interpolates the gain of the noise source between the value specified for the previous frame to that specified for the current value over the duration of the current frame. Interpolation of the gain of the noise source provides a more gradual onset and offset of frication and aspiration. However, when a plosive sound

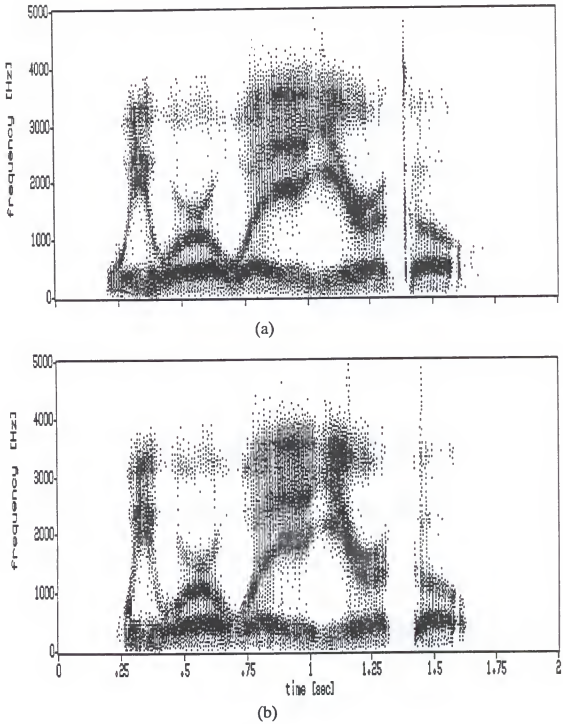


Figure 6–4. Comparison of the spectrograms for the sentence "We were away a year ago": (a) natural speech, and (b) synthetic speech

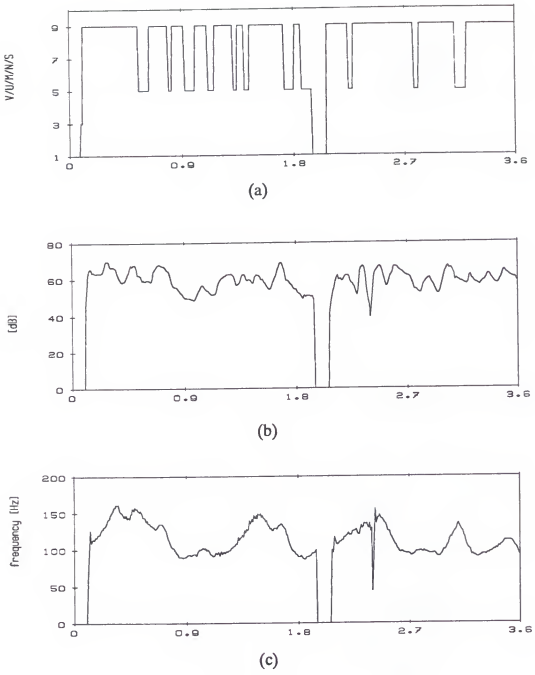
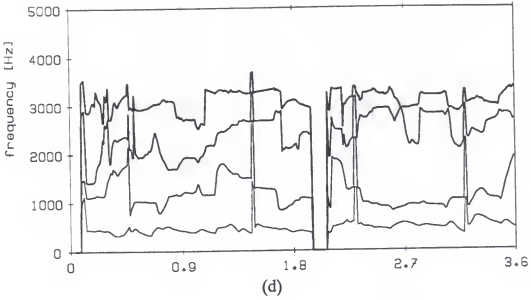
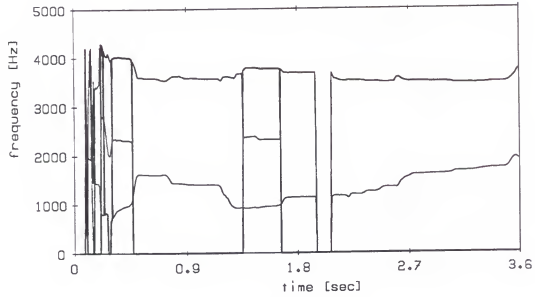


Figure 6–5. Parameter tracks for the sentence "Early one morning a man and a woman ambled along a one mile lane": (a) V/U/M/N/S, (b) gain contoure, (c) fundamental frequency, (d) formant frequency, and (e) antiformant frequency tracks



(d)



(e)

Figure 6–5. Continued

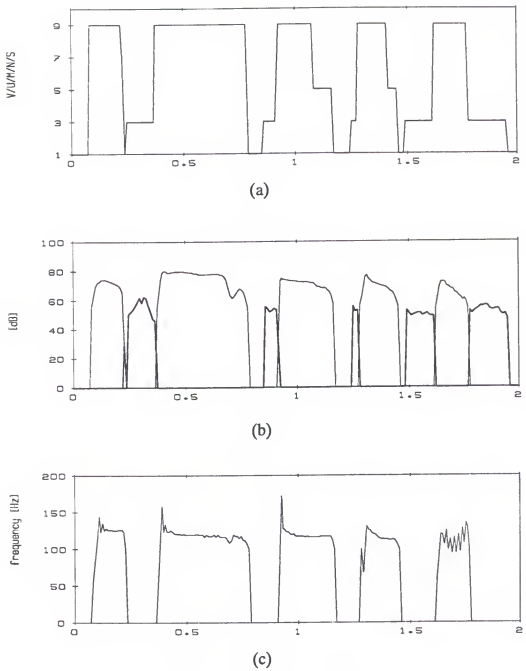
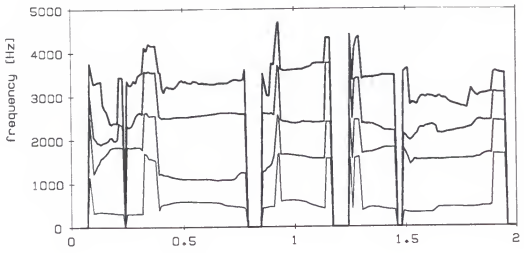
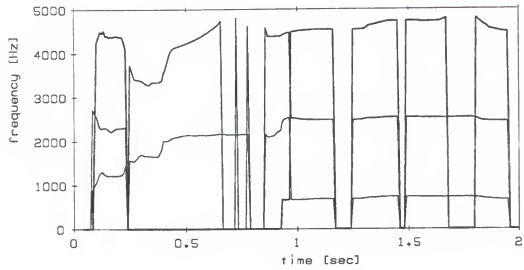


Figure 6–6. Parameter tracks for the sentence "We saw the ten pink fish": (a) V/U/M/N/S, (b) gain contoure, (c) fundamental frequency, (d) formant frequency, and (e) antiformant frequency tracks



(d)



(e)

Figure 6–6. Continued

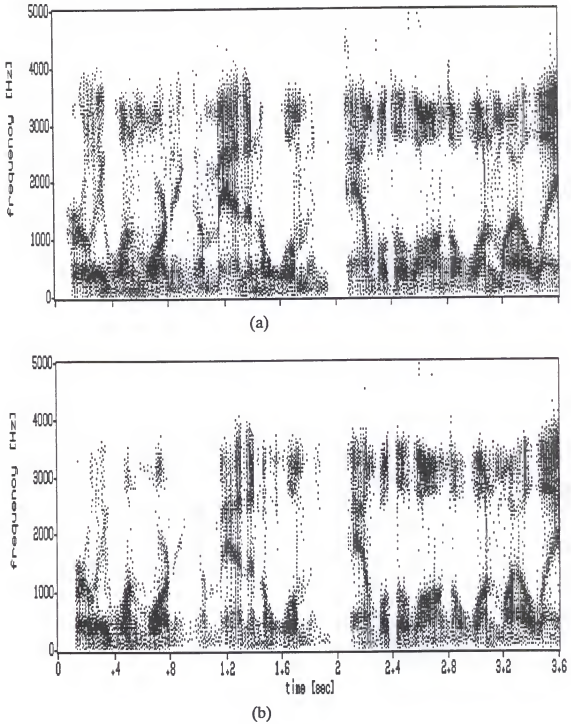


Figure 6–7. Comparison of the spectrograms for the sentence "Early one morning a man and a woman ambled along a one mile lane": (a) natural speech, and (b) synthetic speech

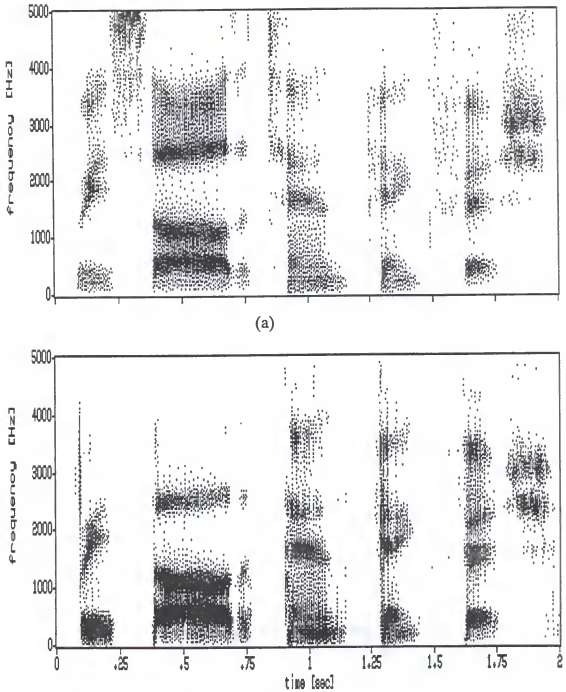


Figure 6–8. Comparison of the spectrograms for the sentence "We saw the ten pink fish": (a) natural speech, and (b) synthetic speech

has to be generated, i.e., when the value of "ah" or the "af" parameter is suddenly increased by 50 dB from its value for the previous frame, the synthesizer increases the gain of the noise source instantaneously to the specified value for the current frame. Klatt (1980) has mentioned a possibility of adding an exponentially decaying pulse to the noise source at the plosive release time in order to simulate the frequency domain aspects of bursts of air flow due to the sudden release of oral pressure behind the plosive occlusion. In the flexible formant synthesizer, the time constant of the exponentially decaying pulse is specified by the "step_size" parameter. The duration of the frication and/or the aspiration noise before the onset of the following voiced sounds is defined as the VOT (Voice Onset Time) of the consonant-vowel syllable. The VOT of the desired value can be obtained by specifying the nonzero values of "ah" or the "af" parameter at least for the duration of VOT and then specifying the nonzero value of both the "f0" and "av" parameters.

The typical values of the formant frequencies, bandwidths and amplitudes given in Klatt (1980) for the fricatives and plosives are valid not only for the frication and aspiration portion of these sounds but also serve as "loci" for the formant trajectories of the consonant-vowel transitions. In the cascade/parallel synthesizer configuration, the aspiration portion is generated through the cascade filter bank since the sound source is located at the glottis. The frication portion is generated by using the parallel filter bank. For fricatives, the energy in the frequency range below the first formant is highly attenuated. Therefore, the first formant generator is not excited by the frication noise source and the amplitude control "a1" is set equal to zero dB. For synthesizing the sibilants (/s/, /ʃ/, /z/ and /ʒ/) a sixth formant generator is used to approximate the high-frequency high-energy level in the spectra of these sounds. The spectra of the fricatives (/f/, /v/, /θ/, /ð/) and the plosives (/p/, /b/) do not show any resonance structure. A by-pass path is used in the parallel filter bank to reproduce the flat spectrum of the noise source for these sounds. Figure 6−9 (a) - (d) show the V/U/M/N/S classification,

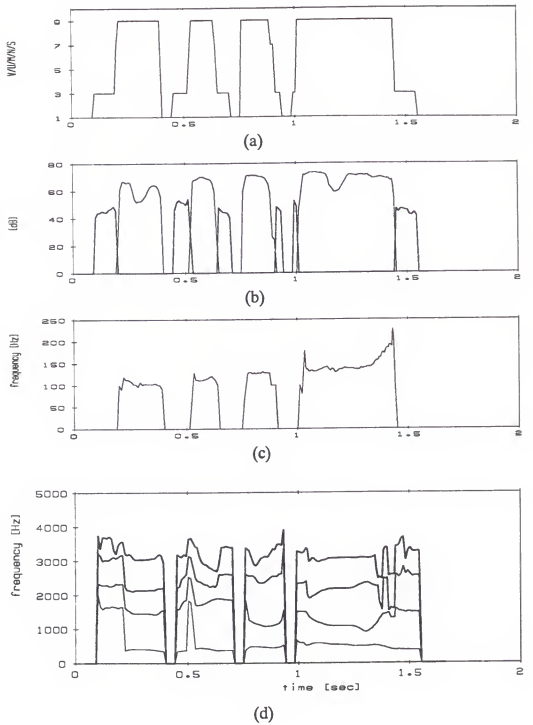


Figure 6–9. Parameter tracks for the sentence "Should we chase those cowboys": (a) V/U/M/N/S, (b) gain contoure, (c) fundamental frequency, and (d) formant frequency tracks

the gain, the fundamental frequency, and the formant frequency contours for the sentence "Should we chase those cowboys?" The spectrogram of natural and synthesized speech utterances of this sentence are shown in Figure 6–10 (a) and (b) respectively. A visual comparison of the two spectrograms demonstrates a good match of the formant frequency tracks and also of the duration of each sound.

6.1.3 Mixed Excitation Sounds (Lalwani, 1991)

During the production of mixed sounds (voiced fricatives and plosives) in the human speech production system, the vibrating vocal folds first modulate the airflow from the lungs and a narrow constriction or an occlusion in the vocal tract then produces a turbulent noise from the modulated air flow. Hence, both the glottal source model and the noise source model have to be used in conjunction when synthesizing mixed excitation sounds. In the flexible formant synthesizer, for synthesizing mixed excitation sounds, the parameters, "f0," "av," and "ah" and/or "af", should be nonzero. The noise source is amplitude modulated pitch-synchronously by an amplitude-time waveform as described earlier.

During the interval before the plosive release there is no sound radiated from the lips. However, there is often a small amount of low-frequency energy radiated from the walls of the throat. This low-frequency energy observed in the spectrograms of voiced plosives is called "voice-bar." When synthesizing the "voice-bar," the synthesized speech consists only of the low-energy glottal source pulses.

In the case of synthesis of voiced fricatives the glottal source pulses are used to excite the cascade filter bank and the frication noise source is used to excite the parallel filter bank. In the case of voiced plosives both the glottal source and the aspiration noise source are used to excite the cascade filter bank and the frication noise source is used to excite the parallel filter bank. Klatt (1980) has pointed out that the advantage of the cascade/parallel synthesizer configuration is that the synthesis of adjacent voiced

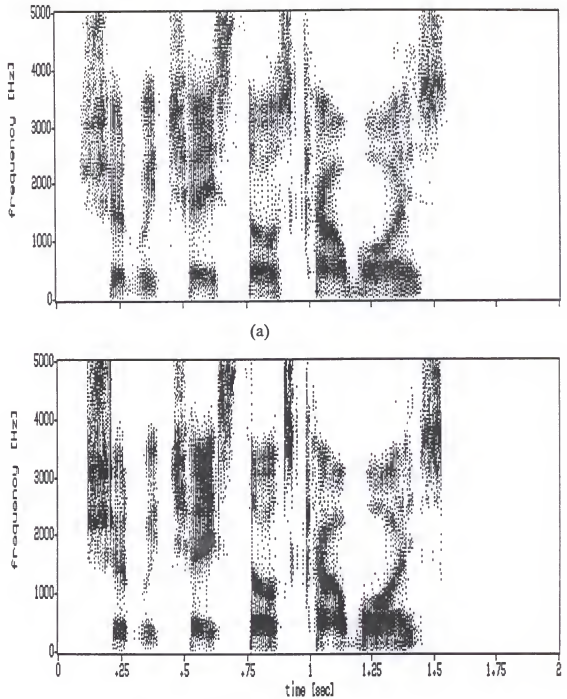


Figure 6–10. Comparison of the spectrograms for the sentence "Should we chase those cowboys": (a) natural speech, and (b) synthetic speech

sounds can be temporally overlapped with the synthesis of frication sounds, such as for syllables /æz/ (as) or /so/ (so), to produce the effect of co-articulation.

When using the all-parallel synthesizer configuration for synthesizing mixed excitation sounds, the first formant generator (resonator) is excited only by glottal source pulses. The second, third, fourth and fifth formant generators (resonators) are excited by a mixture of differentiated glottal pulses and modulated frication (or aspiration) noise. The sixth formant generator and the by-pass path are excited by the frication noise source alone. This strategy, proposed by Klatt (1980) is much simpler than the strategy proposed by Holmes 1983) for producing mixed excitation for formant generators (resonators) in the parallel filter bank. In Figure 6−11 (a) - (e), the V/U/M/N/S classification, the gain contours of the voicing and the fricatives, the fundamental frequency, formant frequency, and antiformant frequency contours for the speech utterance "That zany van is azure." are shown. The spectrograms of the natural and the synthetic speech signal for this utterance are shown in Figure 6−12 (a) and (b). A visual comparison of the two spectrograms demonstrates a good match of the formant antiformant and antiformant frequency tracks and also of the duration of each sound.

6.1.4 Summary

In this chapter we used the flexible formant synthesizer to evaluate our analysis algorithms for several sentences. A visual comparison of the spectrograms of the real and the synthetic speech signals shows a good match for all sentences. We also conducted an informal listening test on the synthsized speech tokens. From this test we found that the quality and the naturalness of the synthesized speech using our analysis algorithms are much better than those using the LPC analysis methods (Lee and Childers, 1988; Childers and Krishnamushy, 1986).

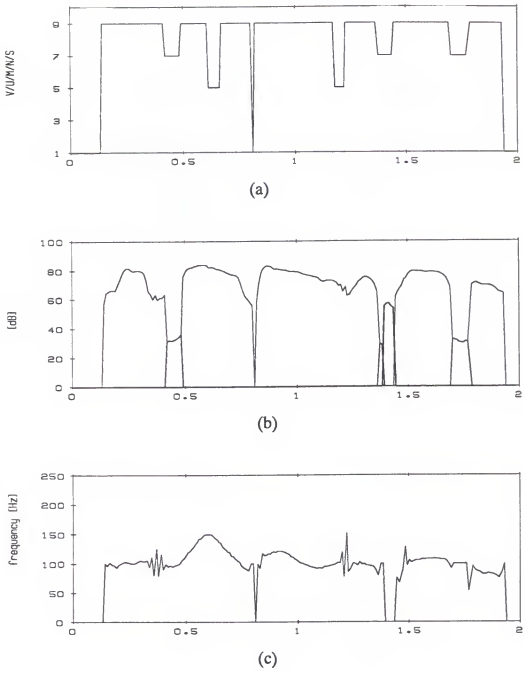
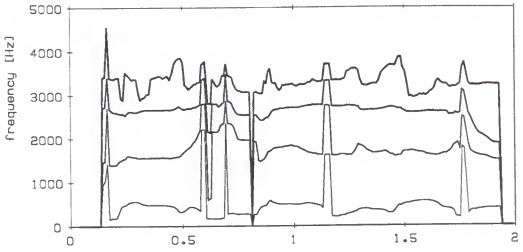
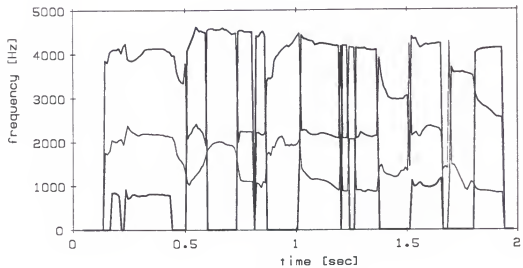


Figure 6–11. Parameter tracks for the sentence "That zany van is azure": (a) V/U/M/N/S, (b) gain contoure, (c) fundamental frequency, (d) formant frequency, and (e) antiformant frequency tracks



(d)



(e)

Figure 6−11. Continued

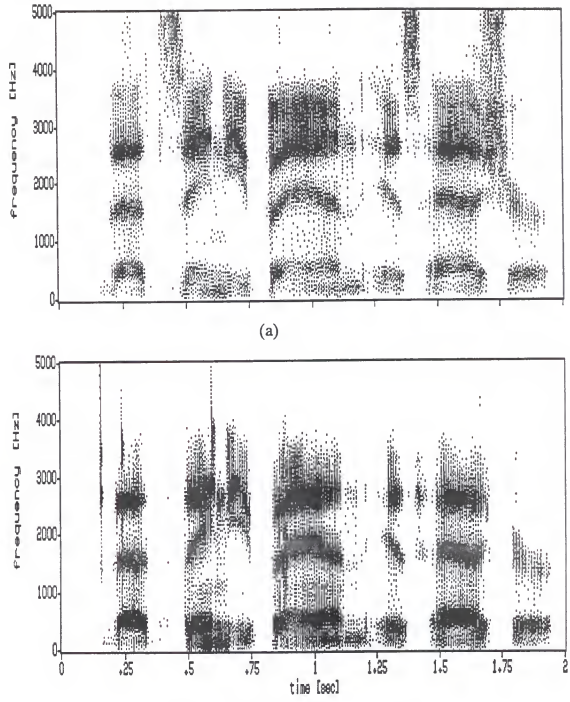


Figure 6–12. Comparison of the spectrograms for the sentence "That zany van is azure": (a) natural speech, and (b) synthetic speech

6.2 Automatic Classification of Different Voice Types

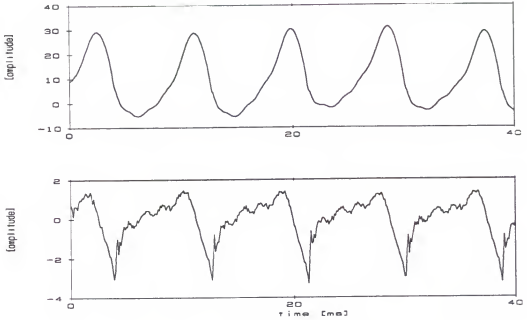
Some glottal waveform characteristics of various voice types are summarized in terms of the glottal factors important for characterizing several voice types (Lee and Childers, 1989; Eskenazi et al., 1990) including: (1) glottal pulse width, (2) glottal pulse skewness, (3) abruptness of glottal closure, (4) turbulent noise, and (5) glottal spectral characteristics. Ahn (1991) described some numerical data measured from inverse filtered glottal waveforms for developing a source model for approximating different voice types.

The objective of this section is to use the LF model parameter sets for automatic classification of an unknown voice type into a known category of voice type using the VQ classifier.

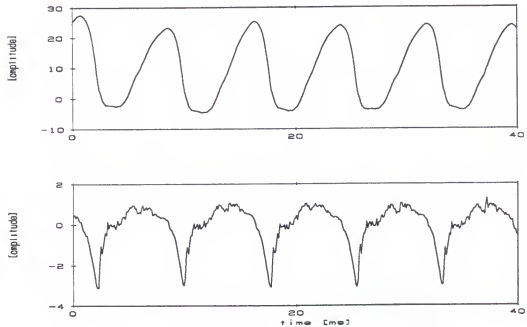
6.2.1 Glottal Waveform Characteristics

Normally a complete vocal fold vibratory cycle (during voiced phonation) consists of an open phase and a closed phase. A wide variation of the glottal waveform shape, its root mean square(rms) intensity and fundamental frequency, phase spectrum, and intensity spectrum have been reported (Sondhi, 1975). The differentiated glottal flow waveforms obtained by inverse filtering the speech signals from different voice types are shown in Figure 6−13. We can see that the glottal flow waveforms are quite different for different voice types.

The glottal flow waveforms may have negative values depending on the starting locations of the analysis frames. The instant of opening is chosen as the starting point for a frame. Since the determination of this instant from the differentiated EGG, LP residual signal, or VFF signal is not very accurate, a frame may start at the middle of the opening phase, and the inverse filtered glottal flow waveform may be negative during the closed phase. This is most noticeable for the vocal fry voice(s), as can be seen

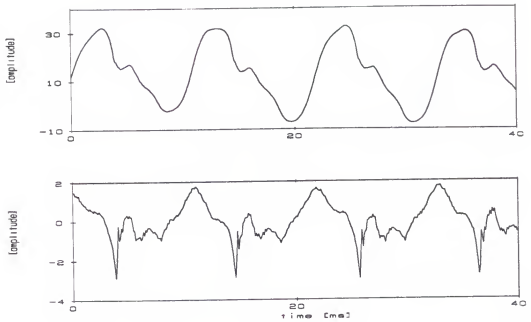


(a) DMH

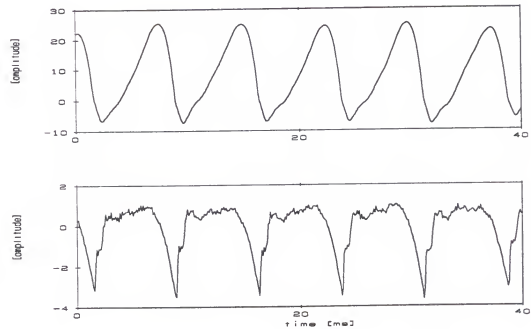


(b) DRW

Figure 6–13. Glottal flow and normalized differentiated glottal flow waveforms for different type voices: (a) & (b) modal voiceds, (c) & (d): vocal fry, (e) & (f): breathy voices

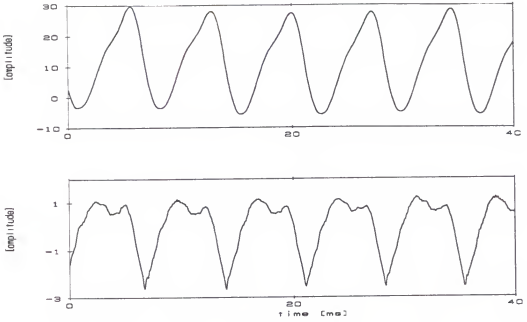


(c) CKL

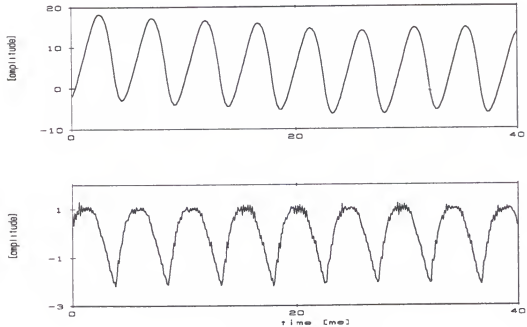


(d) JTO

Figure 6−13. continued



(e) EDR



(f) JMS

Figure 6−13. continued.

in Figure 6−13-(c) and (d). Thus care should be taken in interpreting the opening instants from the inverse filtered differentiated glottal flow waveforms.

The glottal pulse width is moderate for modal voices (Figure 6−13-(a) and (b)) and small for vocal fry voices (Figure 6−13-(c) and (d)). Breathy voices (Figure 6−13-(e) and (f)) have large pulse widths, often making it appear that there is no closed phase. For all the voice types examined (i.e., modal, vocal fry, and breathy voices), the closing phase exhibits a steeper change of slopes than the opening phase. Thus the glottal flow waveforms are skewed to the right. Glottal pulse skewness varies with voice type. For modal and vocal fry phonations the skewing is more apparent than for breathy phonations. Most of the modal and vocal fry voices show very distinct closed phases. The closed phase is not always apparent for breathy voices, and in addition the glottal flow waveforms are somewhat sinusoidal. The glottal closure is relatively steep for modal and vocal fry voices, but progressive for breathy voices.

Due to glottal pulse skewness, the main excitation occurs at the point of vocal fold closure. The magnitude of this excitation can be controlled by the talker over wide ranges (Miller, 1959). In many cases there were also well defined instants of excitation of the second and higher formants at other points in the laryngeal wave (typically these occur at the instant of opening) (Holmes, 1962). For a modal voice, the instant of the maximum closing slope occurs near the instant of glottal closure, resulting in an abrupt termination of the glottal airflow. Vocal fry shows appreciable excitation at the start of the open phase as well as at its end, and there is often an alternation in the spectral content of the excitation from cycle to cycle, causing the relative intensities of the formants to vary (Lee and Childers, 1989; Hunt, 1987). For a breathy voice, the instant of the maximum closing slope occurs near the middle of the glottal closing phase, followed by a residual phase of progressive closure. A breathy voice also shows appreciable formant excitation at the middle of the open phase.

6.2.2 Results of Simple Statistics

The typical mean values and standard deviations (std) of the LF model parameters estimated for different voice types are tabulated in Table 6−1. Note that, in these tables, (1) all mean values and std are expressed in percentage (%) of pitch-period(pp), (2) t_c was computed from the LF model time-function, (3) SQ_{LF} was computed by $SQ_{LF} = t_p/(t_c-t_p)$, and (4) f0 was computed by $f0 = F_s/pp$, where F_s is the sampling frequency (10 kHz).

The statistical results from the data sets reveal clear differences between the mean values of the LF model parameters for different type voices. From this table the mean values of the normalized LF model parameters of different voice types (except the speed quotient and the pitch period) can be ordered in magnitude as fry, modal, and then breathy voice. The mean value of the normalized t_c, which is equivalent to the open quotient (OQ_{LF}) in this study, shows the most notable differences across different type voices. The mean value of the speed quotient (SQ_{LF}) is comparable for modal and vocal fry phonations, but the mean value for a breathy voice is smaller. This means that a breathy voice requires less vocal effort, on the average, than modal or vocal fry voices do. It appears from these data sets that all glottal factors examined from different type voices have continuous numerical values, and the range of these values may overlap for different type voices.

Our data show wide variations of the pitch periods for different type voices: breathy and vocal fry data show comparable variations in pitch period, while modal voice data show a relatively small variation. The ranking of voice type according to increasing pitch period (or, equivalently, decreasing fundamental frequency) was breathy, modal, and vocal fry types, as can be seen from Table 6−1.

6.2.3 Automatic Classification Results

The VQ classifier was used to study the capability of the LF model parameter set to classify an unknown voice type into a specific category. We excluded the parameter pp, the pitch period, from the parameter vector, since the results of the ANOVA tests in Ahn's study (Ahn, 1991) showed this parameter to be of little statistical value. Thus, the model parameter vector consisted of t_p, t_e, t_a, and t_c, and SQ_{LF}.

We divided our data set into two parts: learning data and testing data . Table 6–2 shows the number of the data sets for the VQ classifier. Using these training data we varied the codebook size from one to ten. We calculated the distortion for each codebook size and determined the classification errors for the different voice type classification using the training data. We found that increasing the codebook size beyond eight did not reduce the number of classification errors. We therefore selected our codebook size to be eight using the training data. Thus, we quantized the average distortion measure to eight "levels."

Table 6–1. Mean values and standard deviations (std) of LF model parameters for each subject

Phonation Type	Subject (***)	t_p (%)	t_e (%)	t_a (%)	t_c (%)	SQ_{LF}	pp (ms)	f0 (Hz)
modal	DMHN (431)	43.68 (3.67)	58.26 (5.29)	1.96 (0.96)	67.80 (7.21)	1.95 (0.85)	9.29 (0.90)	108.57 (10.0)
	DRW (463)	36.60 (5.0)	47.67 (6.27)	2.74 (1.09)	60.99 (7.91)	1.70 (1.10)	7.95 (0.69)	126.38 (6.84)
	CKLN (400)	44.32 (3.67)	60.96 (3.31)	2.16 (0.98)	71.53 (5.93)	1.86 (1.26)	8.33 (0.55)	120.5 (7.69)
	average	41.54 (4.28)	55.63 (7.02)	2.28 (0.41)	66.77 (5.34)	1.84 (0.13)	8.52 (0.69)	118.49 (9.07)
vocal fry	CKLP (400)	40.64 (3.47)	55.63 (4.12)	1.46 (0.60)	62.82 (4.01)	1.87 (0.32)	11.37 (0.52)	88.14 (3.93)
	JTO (369)	33.62 (11.45)	44.93 (12.98)	2.05 (1.26)	55.07 (13.80)	1.77 (1.13)	8.17 (1.42)	125.02 (16.37)
	average	37.13 (4.97)	50.28 (7.57)	1.76 (0.29)	58.94 (5.48)	1.82 (0.07)	9.77 (2.26)	106.58 (26.07)
breathy	DMHP (205)	58.31 (5.63)	81.31 (7.62)	2.70 (0.69)	93.17 (4.65)	1.75 (0.46)	10.65 (1.62)	96.27 (17.03)
	EDR (393)	39.59 (6.33)	54.01 (8.44)	4.48 (1.35)	75.01 (11.31)	1.16 (0.34)	7.77 (1.02)	130.64 (15.08)
	GPM (250)	46.71 (11.56)	72.43 (16.28)	2.86 (1.15)	84.56 (12.85)	1.32 (0.85)	9.99 (1.32)	103.31 (35.4)
	JMS (475)	52.27 (7.17)	78.14 (8.94)	4.73 (2.20)	94.38 (7.63)	1.45 (1.10)	5.08 (0.73)	200.36 (34.4)
	average	49.22 (7.98)	71.47 (12.21)	3.69 (1.06)	86.78 (8.98)	1.42 (0.25)	8.37 (2.52)	132.65 (47.5)

* t_p, t_e, t_a, and t_c are in percentage of pitch-period(pp)
* t_c was computed from the LF model
* $SQ_{LF} = t_p/(t_c - t_p)$
* f0 = F_s/pp, where F_s is the sampling frequency(10 kHz)
*** number of data samples used

Table 6−3 and Table 6−4 show the results of the different voice type classifications using the VQ classifier. The classification rates for the learning data sets were 97.14%, 98.5%, and 100% for the modal, vocal fry, and breathy respectively, as can be seen from Table 6−3. The classification rate for the testing data sets were 52.86%, 57.33%, and 73.0% respectively shown in the Table 6−4. According to these tables, it seems that the glottal waveform characteristics of our vocal fry data were relatively similar, on the average, to those for the modal voice. But the vocal fry voice showed more variations in the glottal waveform characteristics than did the modal voice. The testing data sets showed that most of the misclassifications occurred between the modal type and the other voice types. However, misclassifications between the breathy and the vocal fry type were quite small.

Table 6−2. The number of the data sets for the learning and testing of VQ classifier

	modal	vocal fry	breathy	total
learning	700	400	723	1823
testing	594	369	600	1563

In summary, to classify a voice into a certain type based on the glottal factors measured, one should examine the "averaged behavior" of the glottal factors. It can also be said that a modal voice is usually more difficult to correctly identify than the other type voices, while a breathy voice and a fry voice can be differentiated easily. A total of classification rate is 98.5% (1797/1823) for the learning data set and is 61.61% (963/1563) for the testing data set. Therefore, it can be said that the LF model parameter set is a good feature for the classification of the different voice types.

Table 6−3. Classification analysis results for the learning data sets

	Classification results into type		
from type	modal	vocal fry	breathy
modal	680	6	0
vocal fry	20	394	0
breathy	0	0	723
Classificatio n rate (%)	680/700 (97.14 %)	394/400 (98.5 %)	723/723 (100 %)

Table 6−4. Classification analysis results of the testing data sets

	Classification results in % into type		
from type	modal	vocal fry	breathy
modal	314	122	99
vocal fry	189	211	63
breathy	91	36	438
Classificatio n rate (%)	314/594 (52.86 %)	211/369 (57.33 %)	438/600 (73.0 %)

CHAPTER 7
CONCLUSION AND DISCUSSION

7.1 Summary

In this study, the pitch synchronous and asynchronous analysis algorithms were developed for the formant synthesizer.

Three new pitch detection algorithms, which are VFF based, EGG based, and LP error based methods, are introduced. Our pitch detectors are very reliable in quasi-periodic as well as in aperiodic speech signals. The "pitch smearing" effect inherent in speech signal based methods is avoided, and pitch values are available on a period by period basis. Futhermore, the locations of the glottal closing instants allow for the isolation of individual periods of the speech waveform from closure to closure. Within each of these periods, the location of the opening instant is used to further divide the speech signal into closed and open glottis segments. Such fine segmentation of the speech waveform is essential for pitch synchronous analysis algorithms.

A fairly general framework based on a pattern recognition approach to V/U/M/N/S classification has been described in which a set of measurements are made on the interval being classified, and VQ, NN, and decision tree classifiers are used to select the appropriate class. The work constitutes a demonstration that the V/U/M/N/S classification can be made with reasonable accuracy. A VQ classifier achieved 97.5 % classification accuracy on training and 90.85 % accuracy on testing in V/U/M/S classification and achieved 87.02 % on training and 84.41 % for testing in nasal/nonnasal classification. A NN classifier achieved 97.5 % accuracy on training and 96.86% on testing in V/U/M/S classification and achieved 94 % on training and 82.9%

for testing in nasal/nonnasal classification. A decision tree classifier achieved 97.06% in V/U/M/S classification and achieved 82.36% in nasal/nonnasal classification. We have also developed a pitch-synchronous V/U/M/N/S classification method. The method provides the information useful for fine segmentation of the speech signal and also provides the classification rate as good as the pitch-asynchronous algorithm discussed in this study. Pitch synchronous analysis method will be used in a complete analysis/synthesis system.

A new WRLS-VFF-VT algorithm was developed, which is for estimating ARMA parameters, input pulse train, and input white noise at the same time so that formants and antiformants of speech can be correctly estimated. The performance of our algorithm has been tested using synthetic and real speech signals. Our results indicate that the closed phase WRLS-VFF-VT method of analysis seems superior to block processing techniques such as linear predictive, modified Yule-Walker equation methods and other recursive methods such as Kalman filtering, standard WRLS, and weighted least squares lattice (Ting, 1991). The method can be used for all-pole model analysis (e.g., vowels and diphthongs) as well as for pole-zero analysis (e.g., fricatives and nasals). The adaptive recursive algorithms derived from a least square cost function are known to converge rapidly (for short data records) (Haykin, 1985), and have an excellent capability to "track" an unknown parameter vector. In the proposed WRLS-VFF-VT algorithm, a variable forgetting factor is used to allow the estimation process to track the time-varying parameters even more quickly.

A new glottal inverse filtering technique using the WRLS-VFF-VT method was proposed. This method uses the three algorithms for the detection of the glottal closing instants using the VFF, LF error, or EGG signals depending on the characteristics of the input signal. It can provide reliable glottal v-v waveform estimates automatically for the normal speech, synthetic speech, and the pathological speech signals. The proposed algorithm was compared with the two-channel CPC

method for analyzing different types of speech signals. Results showed that the proposed method was capable of estimating the glottal v-v waveform both for the high pitch speech signal and for the speech signal without a completed glottal closed phase (e.g., pathological voices).

The VQ classifier is used to study the capability of the LF model parameter set to classify an unknown voice type into a specific category. The classification rates for the learning data sets are 97.14%, 98.5%, and 100% for the modal, vocal fry, and breathy respectively. The classification rate for the testing data sets are 52.86%, 57.33%, and 73.0%. A total of classification rate is 98.5% (1797/1823) for the learning data set and is 61.61% (963/1563) for the testing data set. Therefore, it can be said that the LF model parameter set is a good feature for the classification of the different voice types.

We use the flexible formant synthesizer to evaluate our analysis algorithms for several sentences. A visual comparison of the spectrograms of the real and the synthetic speech signals shows a good match for all sentences. We also conduct an informal listening test on the synthsized speech tokens. From this test we found that the quality and the naturalness of the synthesized speech using our analysis algorithms are much better than those using the LPC analysis methods (Lee and Childers, 1988; Childers and Krishnamushy, 1986).

7.2 Applications

Our analysis method is useful for both speech synthesis and speech recognition since it identified major acoustic factors of speech signals, especially for the glottal source model. Our complete analysis/synthesis system using our method is being used for voice conversion, synthesis of high-quality speech, and synthesis of various type voices.

7.3 Future Work

The method derived in this study was applied successfully to speech signal analysis. However, there are still some unsolved problems which limit the performance of the proposed algorithm. In general, the problem for selecting the optimal model type and the optimal model order is in most of analysis algorithms. We used the V/U/M/N/S results to select optimal model type for the input speech signals. The problem is the selection of the optimal model order. In this study, the model order is used unchanged value during the process of adaptation. However, by using the V/U/M/N/S results, we can select the optimal order by a statistical value of the model order for each segmentation of speech signals.

The errors of the V/U/M/N/S classification results in the perceptible distortion in the synthesized speech due to the improper selection of the excitation for LPC or using the wrong branch of the formant synthesizer, e.g., cascade instead of parallel. Unfortunately, we didn't consider the nasalized vowels in the nasal/nonnasal detection. So our nasal/nonnasal algorithm needs to be combined with the spectral based algorithm by Yea (Yea and Childers, 1983) for the identification of the nasalized vowels or with another algorithm for the detection of nasalized vowels. Moreover, for the more accurate model type in analysis/synthesis, the more detailed speech segmentation rather than 5-way classification is needed.

A source model was fitted to the glottal flow waveform to obtain the glottal source parameters. The glottal opening instant and the duration of the glottal open phase as well as the closing instant and the closed phase duration were determined from the differentiated EGG signals. However, when we use the speech data only, it is difficult to detect the glottal open phase instants using the LP error, or the VFF signals. We can use the glottal v-v waveform or differentiated glottal v-v waveform to detect

the opening instants directly. So, we need to implement the algorithm for the accurate detection of the opening instants from the glottal v-v waveform.

APPENDIX
DIFFERENT ADAPTIVE FILTERING METHODS FOR SPEECH ANALYSIS

A1 Summary

Methods of estimating time-varying waveform of nonstationary signals using WRLS-VFF (Weighted Recursive Least Square - Variable Forgetting Factor) for ARMA model and ALMS (Adaptive Least Mean Square) for Gamma model and NLMS(Normal Least Mean Square)for AR model are described. For the WRLS-VFF method, the VFF is adapted to a nonstationary signal by an extended prediction error criterion which accounts for the nonstationarity of the signal. This method has a good adaptability in the nonstationary situation and low variance in the stationary situation. For the ALMS method with Gamma model, the filter parameters and μ are adaptive. The Gamma model, the generalized feedforward filter, is a new class of adaptive filters that combine attractive properties of FIR filters with some of the power of IIR filters. ALMS method is of the same computational complexity as the NLMS (Normal Least Mean Square) which has the conventional transversal filter structure. The feasibility of these methods are demonstrated with a speech signal.

A2 Introduction

The purpose of this report is to compare three different adaptive filtering methods, the NLMS filter with ARMA model, the ALMS filter with Gamma model (Principe,

1991) and the WRLS-VFF filter with ARMA model for the application of speech signal processing. Ting (Ting and Childers, 1991) introduced an WRLS-VFF algorithm with ARMA model for speech signal analysis. This algorithm is able to track the time-varying parameters of the vocal tract and update the model parameters during the closed glottal interval, thereby tracking the formants more rapidly and with less error than LPC or other methods. The VFF indicates the state change of the estimator,and can be used to estimate the input excitation when the input is either white noise or periodic pulse trains. In addition,the glottal closed phase interval can be located approximately from the VFF estimation error as we show by comparing the VFF signal with the EGG signal. Thus with the aid of the VFF signal,fully automatic glottal inverse filtering can be achieved. However, computational complexity is a detracting factor with the WRLS-VFF-VT algorithm.

In this study, three different filtering methods will be examined and their performance will be compared in terms of 1) the performance of the estimating time-varying waveform of nonstationary signals, 2) the adaptivity of nonstationary situation, and 3) the computational complexity. Furthermore, we focus on the comparison between the VFF signal in WRLS-VFF and the adaptive μ in ALMS for the application to the closed phase detection and the pitch detection on a period by period basis. Furthermore, we will study whether the Gamma model with ALMS could be used for the speech analysis system or for the speech production system.

A3 System Configuration

A3.1 WRLS-VFF Algorithm for ARMA Model Parameter Estimation

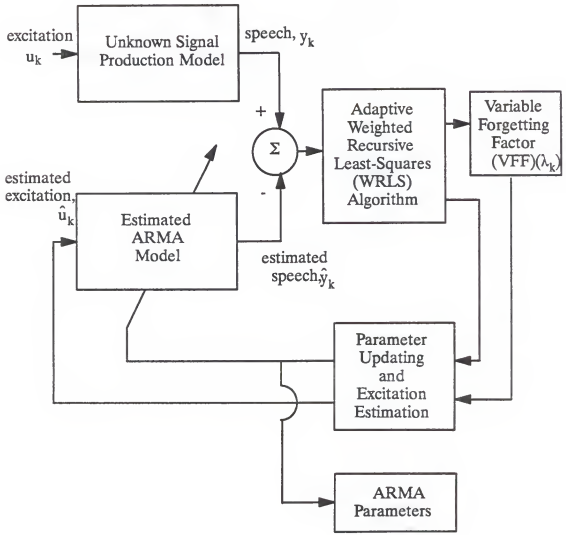


Figure A−1 Block diagram of the WRLS-VFF algorithm for the estimation of ARMA parameter

A3.2 ALMS Algorithm for Gamma Model Parameter Estimation

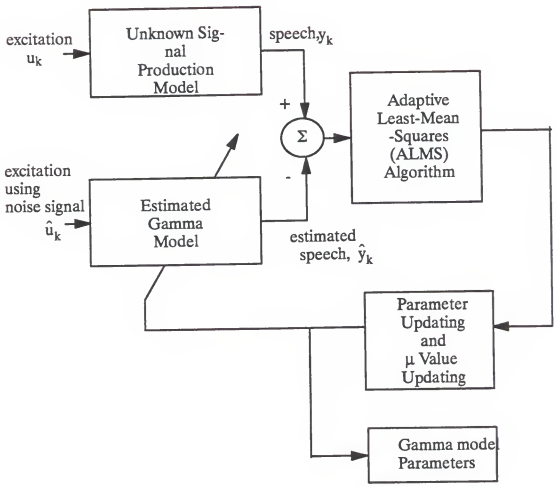


Figure A−2 Block diagram of the ALMS algorithm for the estimation of Gamma model parameter

A4 Analysis of Experiment Results

A4.1 Comparison among DEGG, VFF and Differentiated Adaptive - μ

In Figure A−3, we show the comparison among the DEGG, VFF, and Differentiated Adaptive-μ. The VFF and the differentiated adaptive-μ can be updated recursively at each sample. The VFF indicates the state change of the estimator, and can be used to estimate the input excitation when the input is either white noise or periodic pulse trains. In addition, the glottal closed phase interval can be located approximately from the VFF estimation error as we show by comparing the VFF signal with the EGG signal. The absolute value of the differentiated adaptive-μ has almost the same characteristics as the VFF. For the detection of the pitch period and the input excitation starting point, the VFF and the absolute and differentiated adaptive-μ is useful instead of the EGG signal or the LP residual signal.

A4.2 Comparison of the Estimation Ability among WRLS-VFF with AR Model, ALMS with Gamma Model and NLMS with AR Model

In order to demonstrate the feasibility of AR modeling(8th order)with WRLS-VFF and Gamma modeling(8-th stage)with ALMS, and AR modeling(8-th order)with NLMS, these methods have been applied to nonstationary speech signal ("nine") consisting of nasal and sustained vowel sounds. In Figure A−4, we show the original speech signal and the estimated speech signals. We observe that the estimated signal using the Gamma model with ALMS is slightly better than the one with the AR model with WRLS-VFF and is much better than the one with the AR model with NLMS.

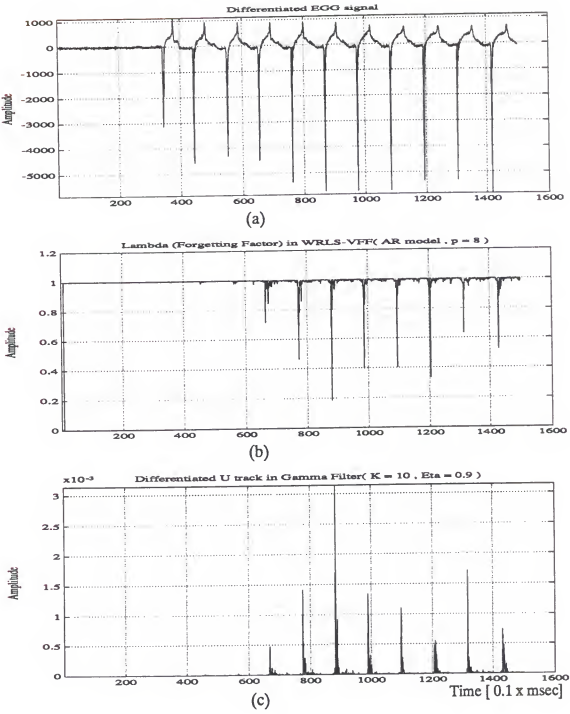


Figure A−3. Comparison among (a) DEGG, (b) VFF, and (c) adaptive μ

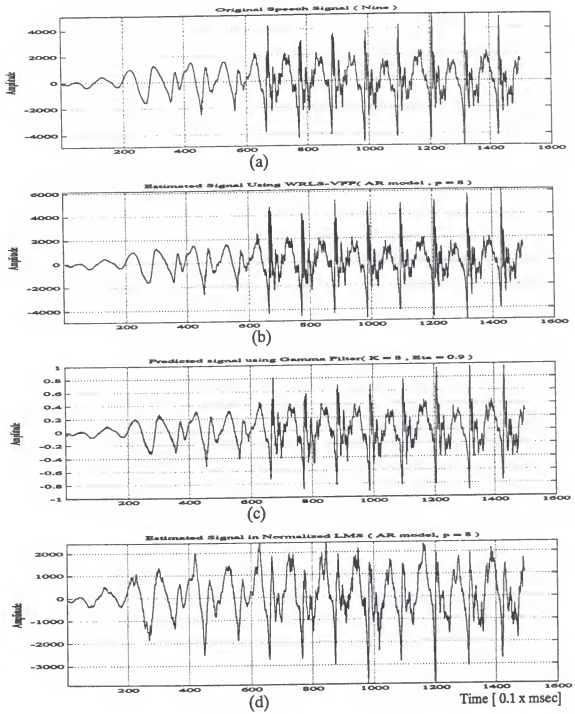


Figure A−4. Comparison of the estimation ability; (a) original signal for a word, "nine", (b) WRLS-VFF with 8-th AR model among WRLS-VFF, (c) ALMS with 8-th gamma model, and (d) NLMS with 8-th AR model

In Figure A−5, we present the estimated errors for the three different methods. The error signal for the AR model with WRLS-VFF is smoother than the one for others in the area between the excitation pulses. However, in the vicinity of the excitation pulses, the error signal for the AR model with WRLS-VFF is rougher than the one for others. The reason is why the adaptive LMS algorithm is known to converge more quickly than the WRLS-VFF algorithm.

In Figure A−6, the comparison for the estimation of a speech signal according to the different model orders in Gamma model with ALMS is shown. There are no big differences among the estimation of speech signals. This means that the low order of the Gamma model can be used to estimate the speech signal. The Gamma model can be used for a new speech production model with low order parameters and with the noise input for the excitation signal.

The characteristics of the individual adaptive μ, accumulated adaptive μ, and the differentiated adaptive μ depending on the different orders of gamma model are shown in Figure A−7, and Figure A−8,Figure A−9 respectively. In these results, the differentiated adaptive μ can be used directly for the detection of the pitch on a period by period basis and for the estimation of the glottal closing instants in speech signals.

A4.3 Comparison of the Computational Complexity among WRLS-VFF with AR Model, ALMS with Gamma Model and NLMS with AR Model

In Table A−1, the computational complexity of each method using the number of flops is shown. The Gamma model with ALMS has a less computational complexity

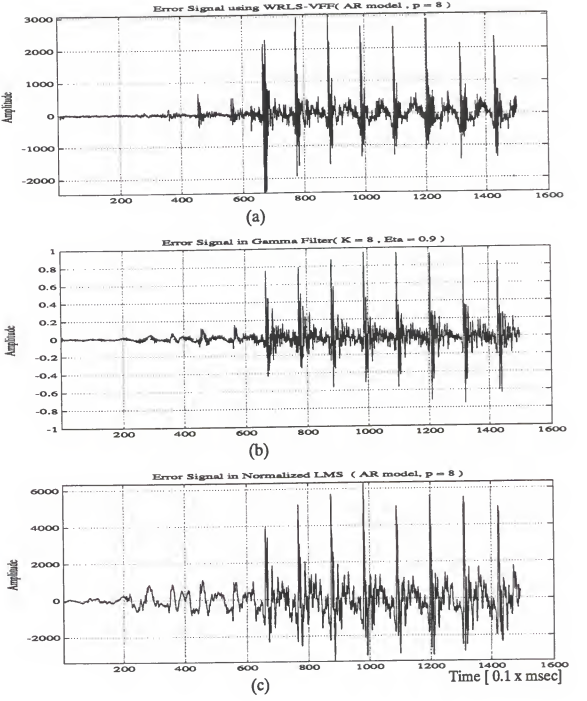


Figure A−5. Comparison of the estimation error signals using; (a) WRLS-VFF with 8-th AR model, (b) ALMS with gamma model, and (c) NLMS with AR model

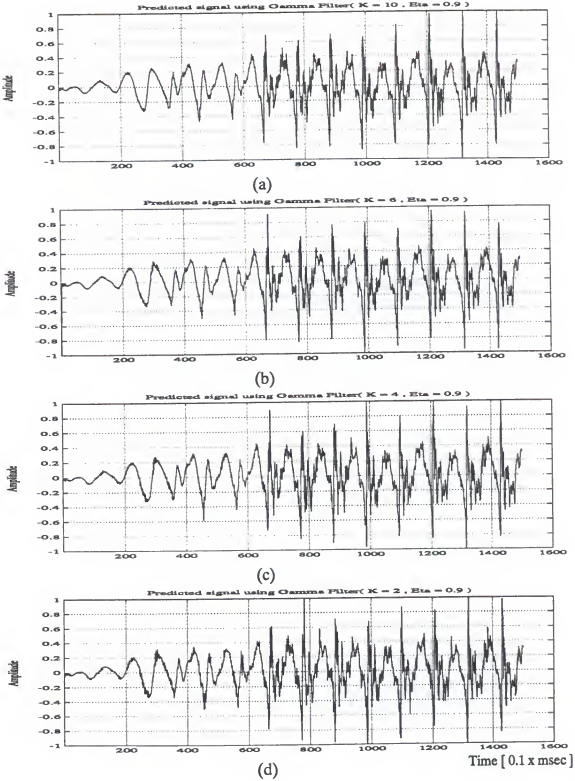


Figure A−6. Comparison of the estimation signal depending on the different orders of gamma model; (a) 10-th order, (b) 6-th order, (c) 4-th order, and (d) 2-nd order

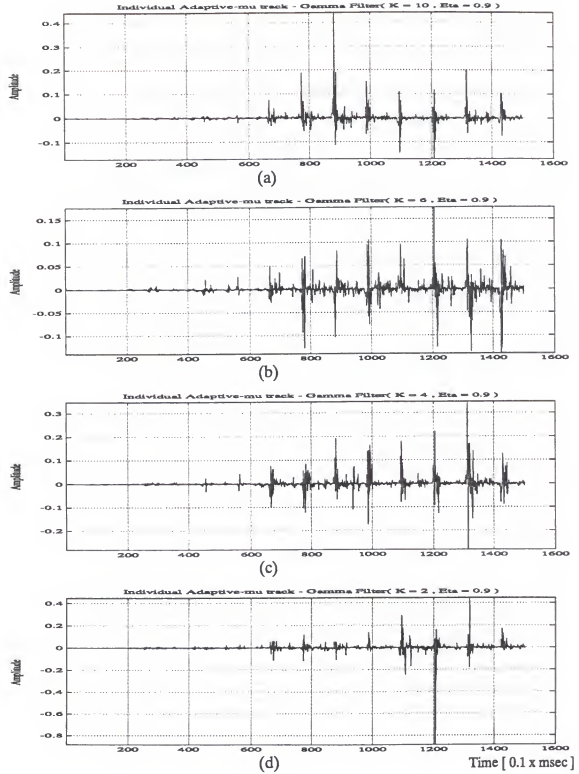


Figure A–7. Characteristics of the individual adaptive μ depending on the different orders of gamma model for: (a) 10-th, (b) 6-th, (c) 4-th, and (d) 2-nd orders

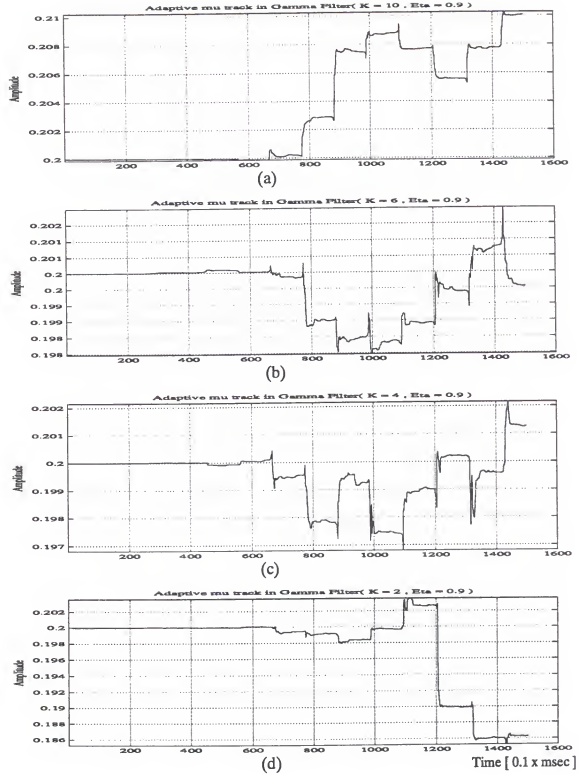


Figure A−8. Characteristics of the accumulated adaptive μ depending on the different orders of gamma model for: (a) 10-th, (b) 6-th, (c) 4-th, and (d) 2-nd orders

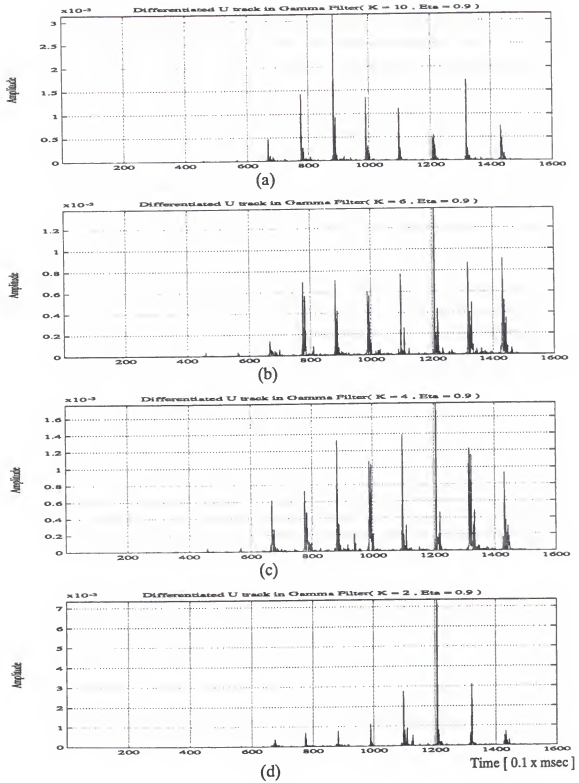


Figure A−9. Characteristics of the absolute and differentiated adaptive-μ depending on the different orders of gamma model for: (a) 10-th, (b) 6-th, (c) 4-th, and (d) 2-nd orders

than the ARMA model with WRLS-VFF algorithm. The definition of the flops is following:

FLOPS Count of floating point operations.

FLOPS returns the cumulative number of floating point operations. It is not feasible to count absolutely all floating point operations, but most of the important ones are counted. Additions and subtractions are one flop if real and two if complex. Multiplications and divisions count one flop each if the result is real and six flops if it is not. Elementary functions count one if real and more if complex. Some examples: If A and B are real N-by-N matrices, then

A + B counts N^2 flops,

A * B counts 2*N^3 flops,

A ^ 100 counts 99*2*N^3 flops,

LU(A) counts roughly (2/3)*N^3 flops.

Table A−1. The Comparison of Computational Complexity for the same filter order, P = 8.

METHODS	NO. OF FLOPS
WRLS-VFF WITH AR MODEL	7,767,808
ALMS WITH GAMMA MODEL	261,173
NLMS WITH AR MODEL	94,124

Table A−2. The number of flops for the different filter orders in Gamma model with ALMS

FILTER ORDER OF GAMMA MODEL	NO. OF FLOPS
2-nd	87,059
4-th	147,097
6-th	201,135
8-th	261,173
10-th	315,211

A5 Conclusion

In this report, an ARMA model with WRLS-VFF and a Gamma model with ALMS for the estimating time-varying waveform of nonstationary speech signals is proposed. The algorithms are able to simultaneously estimate the model parameters at each point as well as the form of the input excitation, including the forgetting factor in the WRLS-VFF and the value of μ in Gamma model. We showed that the VFF and the differentiated adaptive μ can serve as a reliable indicator for the instants of glottal closure by comparing with EGG data.

With respect to the test results, the Gamma model with ALMS is shown to have better adaptability in the nonstationary situation and less computational complexity than the ARMA model with WRLS-VFF.

The Gamma model with ALMS provides a new speech production model with less model parameters and needs only one noise source for source excitation of the model.

From the experimental results, two stages of Gamma filter is a good speech production model with a noise source for input excitation of Gamma model.

According to Figure A−7, Figure A−8, and Figure A−9 the differentiated adaptive μ can be used directly for the detection of the pitch on a period by period basis and the estimation of the glottal closing instants in speech signals.

The Gamma model with ALMS could be used to track the time-varying parameters of the vocal tract and to update the model parameters for the existing speech production such as the AR or ARMA model. This could be done by developing the gamma input-to-output transfer function as the same form as the ARMA model transfer function and by matching the Gamma filter parameter to the ARMA filter parameter. For the this application, there is a limitation in the Gamma model which uses the first order Gamma delay operator, $G(z) = \mu / [z - (1 - \mu)]$. The first order Gamma delay operator will provide the same orders of pole and zero. If the first order Gamma delay operator can be modified to the second order Gamma delay operator, it could be more flexible to apply to the estimation of ARMA model parameters.

The Gamma model with ALMS could be applied to the following areas:

1. Speech analysis-synthesis using Gamma model
2. Speech Coding
3. Speech Communication

REFERENCES

Acronyms:

ASSP - Acoustics, Speech, and Signal Processing
ICASSP - International Conference on Acoustics, Speech, and Signal Processing
IEEE - Institute of Electrical and Electronic Engineers
JASA - Journal of Acoustical Society of America
JSHR - Journal of Speech and Hearing Research
STL-QPSR, RIT - Speech Transmission Lab. Quarterly Progress and Status Report, Royal Inst. of Tech.

Ahn, C. (1991), *A study of voice types and acoustic variabilities: analysis-by-synthesis,* Ph.D. dissertation, University of Florida.

Aikawa, K. and Furui, S. (1988) "Spectral Movement Function and its application to speech recognition," ICASSP-88, N.Y., S5.11, pp. 223-226.

Akaike, H. (1974) "A new look at the statistical model identification," IEEE Trans. Automat. Contr, vol. AC-19, pp. 716-723.

Ananth, A. S., Childers, D. G. and Yegnanarayana, B. (1985) "Measuring source-tract interaction from speech," ICASSP-85, Tampa, FL, pp. 1093-1096.

Andre-Obrech, R. (1988) "A new statistical approach for the automatic segmentation on continuous signals," IEEE Trans. on ASSP, vol. 36, no. 1, pp. 29-40.

Ananthapadmanabha, T. V. (1984) "Acoustic analysis of voice source dynamics," STL-QPSR, RIT, Stockholm, Sweden, vol. 2-3, pp. 1-24.

Ananthapadmanabha, T. V. and Yegnanarayana, B. (1979) "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. on ASSP, vol. 27, pp. 309-319.

Atal, B. S. (1974) "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," JASA, vol. 55, no. 6, pp. 1304-1312 .

Atal, B. S. and Hanauer, S. L. (1971) "Speech analysis and synthesis by Linear Prediction of the speech wave," JASA, vol. 50, pp. 637-655.

Atal, B. S. and Rabiner, L. R. (1976) "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," IEEE Trans. on ASSP, vol. 24, no. 3, pp. 201-212.

Atal, B. S., and Remde, J. S. (1982) "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," Proc. IEEE ICASSP, Paris, France, pp. 614-617.

Atal, B. S., and Schroeder, M. R. (1975) "Recent Advances in Predictive Coding: Applications to Speech Synthesis," Speech Communication (edited by G. Fant, Almgvist and Wiksell), Uppsala, Sweden, vol. 1, pp. 27-31.

Bae, K.S. (1989), *Two-channel analysis: With the application to evaluation of laryngeal function and speaker identification by voice*, Ph.D. dissertation, University of Florida.

Baer, T., Lofqvist, A., and N.S. McGarr (1983) "Laryngeal vibrations: A comparison between high speed filming and glottographic technique," JASA, vol. 73, no. 4, pp. 1304-1308.

Baker, J. M. (1981) "How to achieve recognition: A tutorial/status report on automatic speech recognition," Speech Technology, pp. 30-43.

Bendiksen, Aage and Kenneth Steiglitz (1990) "Neural Networks for Voiced/ Unvoiced Speech Classification," Proceedings of 1990 IEEE ICASSP, Albuquerque, New Mexico, S10.9, pp. 521–524.

Bergem, D. R., Pols, L. C. W. and Beinum, F. J. K. (1988) "Perceptual normalization of the vowels of a man and a child in various contexts," Speech Communication, vol. 7, no. 1, pp. 1-20.

Bergh, A. F., Soong, F. K., and Rabiner, L. R. (1985) "Incorporation of temporal structure into a vector-quantization-based preprocessor for speaker-independent, isolated-word recognition," AT&T Tech. J., vol. 64, no. 5, pp.1047-1063.

Berouti, M. G. (1976) "Estimation of glottal volume-velocity by the linear prediction inverse-filter," Ph.D. dissertation, University of Florida.

Berouti, M. G., Childers, D. G., and Paige, A. (1977) "A correction of tape recorder distortion," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 397-400.

Bickley, C. A. and Stevens, K. N. (1985) "Effects of a vocal tract constriction on the glottal source: Data from voiced consonants," Laryngeal Function in Phonation and Respiration, ed. by T. Baer, C. Sasaki, and K. S. Harris, pp. 239-253, A College-Hill Pub.

Bladon, R. A. W., Henton, C. G., and Pickering, J. B. (1984) "Towards an auditory theory of speaker normalization," Language and Communication, vol. 4, no. 1, pp. 59-69.

Bladon, R. A. W. and Lindblom, B. (1981) "Modeling the judgement of vowel quality difference," JASA, vol. 69, no. 5, pp. 1414-1422.

Blomberg, M., Carlson, R., Elenius, K. and Granstrom, B. (1983) "Auditory models and isolated word recognition," STL-QPSR, RIT, Stockholm, Sweden, vol. 4, pp. 1-15.

Blomberg, M. and Elenius, K. (1986) "Nonlinear frequency warp for speech recognition," ICASSP–86, Tokyo, Japen, vol. 49.2.1, pp. 2631-2634.

Blumstein, S. E. (1986) "On acoustic invariance in speech," Invariance and Variability in Speech Processes, (ed. by J. S. Perkell and D. H. Klatt), Laurence Erlbaum Associates, Hillsdale, NJ, pp. 178-201.

Blumstein, S. E. and Stevens, K. N. (1979) "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, " JASA, vol. 66, no. 4, pp. 1001-1017.

Boone, D. R. (1971), The Voice and Voice Therapy, Prentice-Hall, Englewood Cliffs, NJ.

Brady, P. T. (1968) "A statistical analysis of on-off patterns in 16 conversations," Bell System Technical J., vol. 47, no. 1, pp. 73-91.

Brown, M. K. and Rabiner, L. R. (1982) "On the use of energy in LPC-based recognition of isolated words," Bell System Technical J., vol. 61, pp. 2971-2987.

Buzo, A., Gray, A. H. Jr., Gray, R. M. and Markel, J. D. (1980) "Speech coding based upon vector quantization," IEEE Trans. on ASSP, vol. 28, no. 5, pp. 562-574.

Casscuberta, F. and Vidal, E. (1987) "A nonstationary model for the analysis of transient speech signals," IEEE Trans. on ASSP, vol. 35, pp. 226-228.

Chen, Y. (1988) "Cepstral domain talker stress compensation for robust speech recognition," IEEE Trans. on ASSP, vol. 36, no. 4, pp. 433-439.

Childers, D. G. (1985) "Voice (as opposed to speech) synthesis," Proceedings of the Voice I/O Systems Applications Conference, San Francisco, CA, pp. 350-361.

Childers, D. G. (1987) "Talking computers: Replacing Mel Blanc," Computers in Mech. Eng., pp.22-31, Sep./Oct..

Childers, D. G. and Durling, A. (1975), Digital Filtering and Signal Processing, West Pub. Co., St. Paul.

Childers, D.G, Hahn,M., and J.N. Larar (1989) "Silent and voiced/unvoiced/ mixed excitation (four-way) classification of speech," IEEE Trans. on ASSP, vol. 37, no. 11, pp. 1771-1774.

Childers, D. G., Hicks, D. M., and Moore, G. P. and Alsaka, Y. A. (1986) "A model for vocal fold vibratory motion, contact area, and the electroglottogram," JASA, vol. 80, no. 5, pp. 1309-1320.

Childers, D. G., Hicks, D. M., Moore, G. P., Eskenazi, L., and Lalwani, A. L. (1990) "Electroglottography and vocal fold physiology," JSHR, vol. 33, pp. 245-254 .

Childers, D. G. and Krishnamurthy, A. K. (1985) "A critical review of Electroglottography," CRC Critical Review, Bioengineering, vol. 12, pp. 131-164.

Childers, D. G. and Larar, J. N. (1984) "Electroglottography for laryngeal function assessment and speech analysis," IEEE Trans. on Bio. Eng., vol. 31, no. 12, pp. 807-817.

Childers, D. G. and Lee, C. K. (1991) "Vocal quality factors: Analysis, synthesis, and perception," JASA, vol. 90, no. 5, pp. 2394-2410.

Childers, D. G., Naik, J. M., Larar, J. N., Krishnamurthy, A. K., and Moore, G. P. (1983) "Electroglottography, speech and ultra-high speed cinematography," in Vocal Fold Physiology and Biophysics of Voice, (edited by I.R. Titze and R. Scherer), Denver Center for the Performing Arts, Denver, Chap. 17, pp. 202-220.

Childers, D. G., Skinner, D. P., and Kemerait, R. C. (1977) "The cepstrum: A guide to processing," Proc. of IEEE, vol. 65, no. 10, pp. 1428-1443.

Childers, D. G., Smith, A. M. and Moore, G.P. (1984) "Relationship between electroglottograpy, speech, and vocal contact," Folia Phoniatrica, vol. 36, pp. 105-118.

Childers D. G., and Ting, Y.T. (1991) "Speech analysis using an adaptive weighted recursive least squares algorithm with a variable forgetting factor " submitted for publication.

Childers, D.G. and Wu, K. (1990) "Quality of speech produced by analysis-synthesis," Speech Communication, vol. 9, pp. 97-117.

Childers, D. G., Wu, K., Bae, K. S. and Hicks, D. M. (1988) "Automatic recognition of gender by voice," ICASSP-88, N.Y., pp. 603-606.

Childers, D. G., Wu, K. and Hicks, D. M. (1987a) "Factors in voice quality: Acoustic features related to gender," ICASSP-87, Dallas, Texas, vol. 8.2, pp. 293-296.

Childers, D. G., Wu, K. and Hicks, D. M. (1987b) "Voice conversion: A model for studying voice quality and Speaker normalization," European Conference on Speech Technology, Scotland, pp. 488-491.

Childers, D. G., Wu, K., Hicks, D. M. and Yegnanarayana, B. (1989) "Voice conversion," Speech Communication, vol. 8, pp. 147-158.

Childers, D. G., Yegnanarayana, B. and Wu, K. (1985) "Voice conversion: Factors responsible for quality," ICASSP-85, Tampa, Florida, vol. 19.10, pp. 748-751.

Choukri, K. and Chollet, G. (1986) "Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques," Computer Speech and Language, vol. 1, pp. 95-107.

Cole, A. (ed.) (1980), Perception and Production of Fluent Speech, Lawrence Erlbaum Associates, Hillside, NJ.

Cole, R. A., Stern, R. M. and Larsy, M. J. (1986) "Performing fine phonetic distinctions: Templates versus features," Invariance and Variability in Speech Processes (ed. by J. S. Perkell and D. H. Klatt), Lawrence Erlbaum Associates, Hillside, NJ, pp. 325-345.

Cowan, C.F.N. and P.M. Grant (1985), Adaptive Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Daaboul, F. and Adoul, J. P. (1988) "Parametric segmentation of speech into voiced-unvoiced-silence intervals," Proc. IEEE ICASSP, N.Y., pp. 327-331.

Dautrich, B. A., Rabiner, L. R. and Martin, T. B. (1983) "On the effects of varying filter bank parameters on isolated word recognition," IEEE Trans. on ASSP, vol. 31, no. 4, pp. 793-806.

Deller, Jr. J. R. (1981) "Some notes on closed phase glottal inverse filtering," IEEE Trans. on ASSP, vol. 29, no. 4, pp. 917-919.

Demichelis P., (1982) "The automatic recognition of some sonorant consonants in continuous speech," Speech Communication, vol. 1, pp. 231-255.

Doddington, G. R. (1980) "Whither speech recognition?," Trends in Speech Recognition (ed. by W. Lea), NJ, Prentice Hall, Englewood Cliffs, pp. 556-561.

Edwards, A. L. (1973), Statistical Methods, 3rd ed., Holt, Rinehart and Winston, New York.

Eskenazi, L., Childers, D. G., Hicks, D. M. (1990) "Acoustic correlates of vocal quality," JSHR, vol. 33, pp. 298-306 .

Fant, G. (1960), Acoustic Theory of Speech Production, Mouton, The Hague.

Fant, G. (1979) "Glottal source and excitation analysis," STL-QPSR, RIT, Stockholm, Sweden, vol. 1, pp. 85-107.

Fant, G. (1980) "Voice source dynamics," STL-QPSR, RIT, Stockholm, Sweden, vol. 2-3, pp. 17-37.

Fant, G. and Ananthapadmanabha, T. V. (1982) "Truncation an superposition," STL-QPSR, RIT, Stockholm, Sweden, vol. 2-3, pp. 1-17.

Fant, G., Liljencrants, J., and Lin, Q. G. (1985) "A four-parameter model of glottal flow," STL-QPSR, RIT, Stockholm, Sweden, vol. 4, pp. 1-13.

Fant, G. and Lin, Q. G. (1988) "Frequency domain interpretation and derivation of glottal flow parameters," STL-QPSR, RIT, Stockholm, Sweden, vol. 2-3, pp. 1-21.

Flanagan, J. L. (1972), Speech Analysis, Synthesis and Perception, 2nd ed. Springer-Verlag, New York.

Flanagan, J. L., Ishizaka, K., and Shipley, K. L. (1975) "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell System Technical Journal, vol. 54, pp. 485-506.

Fortescue, T.R., Kershenbaum, L.S. and Ydstie, B.E. ; "Implementation of self-regulators with variable forgetting factors," Automatica, vol. 17, pp. 831-835, 1981.

Forrest, K., Weismer. G., Milenkovic, P., and Dougall, R. (1988) "Statistical analysis of word-initial voiceless obstruents: Preliminary data," JASA, vol. 84, pp. 115-123.

Friedlander, B. (1982) "A recursive maximum likelihood algorithm for ARMA spectral estimation," IEEE Trans. Information Theory, vol. IT-28, no. 4, pp. 639-646.

Frokjaer-Jensen, B. and Prytz, S. (1976), "Registration of voice quality," Bruel and Kjaer Technical Review, vol. 3, pp. 3-17.

Fujimura, O. (1960) "Spectra of Nasalized Vowels," Res. Lab. Electron. Q. Prog. Rep. (MIT, Cambridge, MA), vol. 58, pp. 214-218.

Fujimura, O. (1961) "Analysis of Nasalized Vowels," Res. Lab. Electron. Q. Prog. Rep. (MIT, Cambridge, MA) vol. 62, pp. 191-192.

Fujimura, O. (1962) "Analysis of nasal consonants," JASA, vol. 34, pp. 1865-1875.

Fujisaki, H. and Ljungqvist, M. (1986), "Proposal and evaluation of models for the glottal source waveform," Proc. IEEE ICASSP, Tokyo, Japan, pp. 1605-1608.

Furui, S. (1981) "Cepstral analysis techniques for automatic speaker verification," IEEE Trans. on ASSP, vol. 29, no. 2, pp. 254-272.

Furui, S. (1986a) "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. on ASSP, vol. 34, no. 1, pp. 52-59.

Furui, S. (1986b) "On the role of spectral transition for speech perception," JASA, vol. 80, no. 4, pp. 1016-1025.

Furui, S. (1988) "A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition," IEEE Trans. on ASSP, vol. 36, no. 7, pp. 980-987.

Gray, R. M. (1984) "Vector quantization," IEEE ASSP Magazine, pp. 4-29.

Gray, R. M., Buzo, A., Gray, A. H. Jr. and Matsuyama, Y. (1980) "Distortion measures for speech processing," IEEE Trans. on ASSP, vol. 28, pp. 367-376.

Gray, Jr. A. H. and Markel, J. D. (1976) "Distance measures for speech processing," IEEE Trans. on ASSP, vol. 24, no. 5, pp. 380-391.

Grenier, Y. (1983) "Time-dependent ARMA modeling of nonstationary signals," IEEE Trans. on ASSP, vol. 31, pp. 899-911.

Guerin, B., Mrayati, M., and Carre, R. (1976) "A voice source taking account of coupling with the supraglottal cavities," IEEE ICASSP, Philadelphia, Pennsylvania, vol. 1, pp. 47-50.

Gupta, V. N., Bryan, J. K. and Gowdy, J. N. (1978) "A speakerindependent speech-recognition system based on linear prediction," IEEE Trans. on ASSP, vol. 26, no. 1, pp. 27-33.

Hall, M. G., Oppenheim A. V., and Willsky, A. S. (1983) "Time-varying parameteric modeling of speech," Signal Processing, vol. 5, pp. 267-285.

Hahn, M (1989), *Silence and voice-unvoiced-mixed excitation classification of speech with applications: A two-channel and a one-channel,* Ph.D. dissertation, University of Florida.

Hanson, B. A. and Wakita, H. (1987) "Spectral slope distance measures with linear prediction analysis for word recognition in noise," IEEE Trans. on ASSP, vol. 35, no. 7, pp. 968-973.

Haykin, S. (1985) Adaptive Filter Theory Prentice-Hall, Inc., Englewood Cliffs, NJ.

Hecker, M. (1971) "Speaker recognition: An interpretive survey of the literature," no. 16, Monograph of America Soc. of Hearing Aid.

Hedelin, P. (1984) "A glottal LPC-vocoder," ICASSP-84, San diego, California, pp. 1.6.1-1.6.4.

Hedelin, P. (1986) "High quality glottal LPC-vocoding," ICASSP-86, Tokyo, Japen, vol. 9.9, pp. 465-468.

Hermansky, H. (1987) "An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception," ICASSP-87, Dallas, Texas, vol. 27.10, pp. 1159-1162.

Hermansky, H. and Junqua, J. C. (1988) "Optimization of perceptually based ASR front end," ICASSP-88, N.Y., S5.10, pp. 219-222.

Hess, W. J. (1982) "Algorithms and Devices for Pitch Determination of Speech Signals," Phonetica, vol. 39, pp. 219-240.

Hillman, R. E., Oesterle, E. and Feth, L. L. (1983) "Characteristics of the glottal turbulent noise source," JASA, vol. 74, pp. 691-694.

Hillman, R. E. and Weinberg, B. (1981) "Estimation of glottal volume velocity waveform properties: A review and study of some methodological assumptions," Speech and language: Advances in Basic Research and Practice (ed. by N. Lass), Academic Press, New York, vol. 6, pp. 411-473.

Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S. and Tanabe, M. (1984) "Harmonic-intensity analysis of normal and hoarse voices," JASA, vol. 76, pp. 1648-1651.

Hollien, H (1974), "On vocal register," Journal of Phonetics, vol. 2, pp. 125-144.

Holmberg, E. B., Hillman, R. E. and Perkell, J. S. (1988) "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," JASA, vol. 84, no. 2, pp. 511-529.

Holmes, J. N. (1962) "An investigation of the volume velocity waveform at the larynx during speech by means of an inverse filter," Proc. IV Internat. Congress on Acoustics, pp. 1-4.

Holmes, J. N. (1973) "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE Trans. Audio and Electroacoustics, vol. 21, pp. 298-305 .

Holmes, J.N. (1976) "Formant excitation before and after glottal closure," IEEE ICASSP 1976, Philadelphia, Pennsylvania, pp. 39-42.

Holmes, J. N. (1983) "Formant synthesizers: cascade or parallel?" Speech Communication, vol. 2, pp. 251-273.

Holmes, J. N. (1986) "Normalization in vowel perception," Invariance and Variability in Speech Processes, ed. by J. S. Perkell and D. H. Klatt, Lawrence Erlbaum Associates, Hillside, NJ, pp. 346-359.

Holm, S. (1981) "Automatic generation of mixed excitation in a Linear Predictive speech synthesizer," Proc. IEEE ICASSP, Atlanta, GA, pp. 118-120.

Hunt, M. J., Bridle, J. S. and Holmes, J. N. (1978) "Interactive digital inverse filtering and its relation to linear prediction methods," IEEE ICASSP, Tulsa, Oklahoma, pp. 15-18.

Ishizaka, K. and Flanagan, J. L. (1972) "Synthesis of voiced sounds from a two-mass model of the vocal cords," Bell System Technical Journal, vol. 51, pp. 1233-1268.

Itakura, F. (1975) "Minimum prediction residual principle applied to speech recognition," IEEE Trans. on ASSP, vol. 23, pp. 67-72.

Itakura, F., and Saito, S. (1968) "Analysis-synthesis telephony based upon the masimum likelihood method," Proceedings of the 6th International Congress on Acoustics, pp. 393-400.

Juang, B. H. (1984a) "On the hidden Markov model and dynamic time warping for speech recognition - A unified view," AT&T Bell Lab. Technical J., vol. 63, no. 7, pp. 1213-1243.

Juang, B. H. (1984b) "On using the Itakura-Saito measures for speech coder performance evaluation," AT&T Bell Lab. Technical J., vol. 63, no. 8, pp. 1477-1498.

Juang, B. H., Rabiner, L. R. and Wilpon, J. G. (1987) "On the use of bandpass liftering in speech recognition," IEEE Trans. on ASSP, vol. 35, no. 7, pp. 947-954.

Kay, S. (1987) Modern Spectrum Estimation, Prentice-Hall, Inc. Englewood Cliffs, NJ.

Kaneko, T. and Dixon, R. (1983) "A hierarchical decision approach to large vocabulary discrete utterance recognition," IEEE Trans. on ASSP, vol. 31, pp. 1061-1066.

Keeler, L. O., Clement, G. L., Strong, W. J., and Palmer, E. P. (1976) "Two preliminary studies of the intelligibility of predictorcoefficient and formant-coded speech," IEEE Trans. on ASSP, vol. 24, pp. 429-432.

Kewley-Port, D. (1983) "Time-varying features as correlates of place of articulation in stop consonants," JASA, vol. 73, no. 1, pp. 322-335.

Klatt, D. H. (1977) "Review of the ARPA speech understanding project," JASA, vol. 62, pp. 1345-1366.

Klatt, D. H. (1980) "Software for a cascade/parallel formant synthesizer," JASA, vol. 67, no. 3, pp. 971-995.

Klatt, D. H. (1980) "SCRIBER and LAFS: two new approach to speech analysis," Trends in Speech Recognition (ed. by W. Lea),NJ, Prentice Hall, Englewood Cliffs, pp. 529-555.

Klatt, D. H. (1982) "Prediction of perceived phonetic distance from critical-band spectra: A first step," ICASSP-82, Paris, France, pp. 1278-1281.

Klatt, D. H. (1986) "The problem of variability in speech recognition and in models of speech perception," Invariance and Variability in Speech Processes (ed. by J. S. Perkell and D. H. Klatt), Lawrence Erlbaum Associates, Hillside, NJ, pp. 300-324.

Klatt, D. H. (1987) "Review of text-to-speech conversion for English," JASA, vol. 82, pp. 737-793.

Klatt, D. H. and Klatt, L. C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers," JASA, vol. 87, no. 2, pp. 820-857.

Krishnamurthy, A. K. and Childers, D. G. (1986) "Two-channel speech analysis," IEEE Trans. on ASSP, vol. 34, no. 4, pp. 730-743.

Kurowski, K. and Blumsten, S.E.(1984) "Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants," JASA, vol. 76, pp. 383-390.

Kurowski, K. and Blumstein, S.E.(1987) "Acoustic properties for place of articulation in nasal consonants," JASA, vol. 81, pp. 1917-1927.

Kuwabara, H. (1985) "An approach to normalization of coarticulation effects for vowels in connected speech," JASA, vol. 77, no. 2, pp. 686-694.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G. and Scherer, K. R. (1985) "Evidence for the independent function of intonation contour type voice quality, and F0 range in signaling speaker affect," JASA, vol. 78, no. 2, pp. 435-444ust.

Ladefoged, P., Kameny, I. and Brackenbridge, W. (1976) "Acoustic effect of style," JASA, vol. 59, no. 1, pp. 228-231.

Lahiri, A. Gewirth, L. and Blumstein, S. E. (1984) "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," JASA, vol 76, pp. 391-404.

Lalwani, A. L. and Childers D. G. (1991) "Modeling vocal disorders via formant synthesis," ICASSP, Ontario, Canada, pp. 505-508.

Landau, I.D. (1976) "Unbiased recursive identification using model reference adaptive technique," IEEE Trans. Automat. Contr., vol. AC-21, pp. 194-202.

Larar, J. N. (1985). *Towards speaker independent isolated word recognition for large lexicons: A two-channel, two-pass approach*, Ph.D. dissertation, University of Florida.

Larar, J. N. (1986) "Lexical access using acoustic-phonetic classifications," Computer Speech and Language I, Academic Press, New York, pp. 47-59.

Larar, J. N., Alsaka, Y. A. and Childers, D. G. (1985) "Variability in closed phase analysis of speech," ICASSP-85, Tampa, Florida, pp. 1089-1092.

Laver, J. (1980), The Phonetic Description of Voice Quality, Cambridge University Press, New York.

Laver, J. and Hanson, R. (1981) "Describing the normal voice," Evaluation of Speech in Psychiatry (ed. by J. Darby), Grune and Stratton, New York, pp. 51-78.

Lea, W. A. (1980) "Speech recognition: past, present, and future," Trends in Speech Recognition, ed. by W. Lea, Prentice Hall, Englewood Cliffs, NJ, pp. 39-98.

Lea, W. A. (1983) "Selecting the best speech recognizer for the job," Speech Technology, pp. 10-29.

Lebrum, Y. (1971) "On the so called dissociations between electroglottogram and phonogram," Folia Phoniatrica, vol. 23, pp. 225-227.

Lee, C. K. (1988) *Voice quality: Analysis and synthesis*, Ph.D. dissertation, University of Florida.

Lee, C. K., Childers, D. G. (1989) "Some acoustical, perceptual, and physiological aspects of vocal quality," Vocal Fold Physiology Conference (to appear in book form, Raven Press), Stockholm, Sweden, pp. 1–8, July.

Lee, D.T.L. and Morf, M. and Friedlander, B. (1981) "Recursive least squares ladder estimation algorithms," IEEE Trans. on ASSP, vol. 29, no. 3, pp. 627-641.

Levinson, S. E. (1985) "Structural methods in automatic speech recognition," IEEE Proc., vol. 73, no. 11, pp. 1625-1650.

Levinson, S. E. (1987) "Statistical methods for speaker independence," Fundamentals in Computer Understanding: Speech and Vision (ed. by J. P. Haton) University Press, Great Britain, Cambridge, pp. 207-216.

Lieberman, P. (1961) "Perturbation in vocal pitch," JASA, vol. 33, pp. 597-603.

Lieberman, P. (1963) "Some acoustic measures of the fundamental periodicity of normal and pathological larynges," JASA, vol. 35, pp. 344-353.

Linde, Y. L., Buzo, A. and Gray, R. M. (1980) "An algorithm for vector quantizer design," IEEE Trans. on Communication, vol. 28, no. 1, pp. 84-95.

Liporace, L. A. (1975) "Linear estimation of nonstationary signals," JASA, vol. 58, pp. 1288-1295.

Lippmann, R. P. (1988) "Neural nets for computing," ICASSP-88, N.Y., pp. 1-6.

Lippmann, R. P., Mack, M. M. and Paul, D. P. (1986) "Multi-style training for robust speech recognition under stress," JASA, Suppl. 1, vol. 79, pp. S95.

Makhoul, J. (1976) "Linear prediction: A tutorial review," Proc. IEEE, vol. 64, pp. 99-118.

Makhoul, J., Roucos S. and Gish H. (1985) "Vector quantization in speech coding," IEEE Proc., vol. 73, no. 11, pp. 1551-1588.

Makino, S. and Kido K. (1986) "Recognition of phonemes using timespectrum pattern," Speech Communication, vol. 5, pp. 225-237.

Markel, J.D. (1972) "The SIFT algorithm for fundamental frequency estimation," IEEE Trans., vol. AU-20, pp. 367-377.

Markel, J.D. (1973) "Basic formant and F0 parameter extraction from a digital inverse filter formulation," IEEE Trans., vol. AU-21, pp. 154-160.

Markel, J. D. and Gray, A. H. (1976), Linear Prediction of Speech, Springer-Verlag, New York.

Marple, S.L. (1987) Digital Spectral Analysis with Applications, Prentice-Hall, Inc., Englewood Cliffs, NJ.

Martin, S. L. (1987) "Speech Technology finding real application," Computer Design, pp. 38-43, Mar.

Matausek, M.R. and Batalov, V.S. (1980) "A new approach to the determination of the glottal waveform," IEEE Trans. on ASSP, vol. 28, pp. 616-622.

Mathews, M. V., Miller, J. E., and David, E. E. (1961) "Pitch synchronous analysis of voiced sounds," JASA. vol. 33, pp. 179-186.

Matsumoto, H. and Wakita, H. (1986) "Vowel normalization by frequency warped spectral matching," Speech Communication, vol. 5, pp. 239-251.

McCandless, S. S. (1974) "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. on ASSP, vol. 22, pp. 135-141.

Mermelstein, P. (1977) "On detecting nasals in continuous speech," JASA, vol. 61, pp. 581-587.

Milenkovic, P. (1986) "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. on ASSP, vol. 34, no. 1, pp. 28-41.

Milenkovic, P. and Mo, F. (1986) "Glottal inverse filtering of nasalized vowels," JASA, Suppl. 1, vol. 80, pp. S19.

Miller, J. D. (1959) "Nature of the vocal cord wave," JASA, vol. 31, pp. 667-677 .

Miller, J. D., Engebretson, A. M. and Vemula, N. R. (1980) "Vowel normalization: Differences between vowels spoken by children, women, and men," JASA, vol. 68, suppl. 1, pp. 533.

Milenkovic, P. (1986) "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Trans. on ASSP, vol. 34, no. 1, pp. 28-42.

Minsoo, H. (1990) *Silence and voiced - unvoiced - mixed excitation classification of speech with applications: a two- channel and a one- channel*, Ph.D dissertation, University of Florida.

Miyanaga, Y.,Miki, N.,Nagai, N. and Hatori, K. (1982) "A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system," IEEE Trans. on ASSP, vol. 30, no. 1, 88-95.

Miyanaga, Y.,Miki, N. and Nagai, N. (1986) "Adaptive identification of a time-varying ARMA speech model," IEEE Trans. on ASSP, vol. 34, pp. 423-433.

Miyoshi, Y., Yamato, K., Mizoguchi, R., Yanagida, M. and Kakusho, O. (1987) "Analysis of speech signals of short pitch period by a sampleselective linear prediction," IEEE Trans. on ASSP, vol. 35, no. 9, pp. 1233-1239.

Monsen, R. B. and Engebretson, A. M. (1977) "Study of variations in the male and female glottal wave," JASA, vol. 62, no. 4, pp. 981-993.

Moore, P. (1962), "Observation on the physiology of hoarseness," Proc. 4th Int. Congress of Phonetic Sci., Helsinki, Filand, pp. 92-95.

Mori, R. D., Gubrynowicz, R., and Laface, P. (1979) "Inference of knowledge source for the recognition of nasals in continuous speech," IEEE Trans. on ASSP, vol. 27, no. 5, pp. 538-549.

Mori, R. D., Lam, L. and Probst, C. (1987) "Rule-based detection of speech features for automatic speech recognition," Fundamentals in Computer Understanding: Speech and Vision (ed. by J. P. Haton), Cambridge University Press, Cambridge, pp. 155-179.

Morikawa, H. and Fujisaki, H. (1982) "Adaptive analysis of speech based on a pole-zero representation," IEEE Trans. on ASSP, vol. 30, no. 1, pp. 77-87.

Morikawa H. and Fujisaki, H.(1984) "System identification of the speech production process based on a state-space representation," IEEE Trans. on ASSP, vol. 32, no. 2, pp. 252-262.

Nadas, A., Nahamoo, D. and Picheny, M. A. (1988) "On a model-robust training method for speech recognition," IEEE Trans. on ASSP, vol. 36, no. 9, pp. 1432-1435.

Naik, J. M. (1983) *Synthesis and evaluation of natural sounding speech using the linear predictive analysis-synthesis scheme*, Ph.D. dissertation, University of Florida.

Nakata, K. (1959) "Synthesis and perception of nasal consonants," JASA, vol. 31, pp. 661-666.

Neuburg, E. P. (1979), "Automatic thresholding for voicing detection algorithms," Proc. IEEE ICASSP, Washington D.C., pp. 756-758.

Nocerino, N., Soong, F. K., Rabiner, L. R. and Klatt, D. H. (1985) "Comparative study of several distortion measures for speech recognition," Speech Communication, vol. 4, pp. 317-331.

O'Shaughnessy,D. (1983) "Automatic speech synthesis," IEEE Communications Magazine, vol. 21, no. 12, pp. 26-34.

Pagano, M. (1974) "Estimation of autoregressive signals plus white noise," Ann. Statist., vol. 2, no. 1, pp. 97-108.

Paliwal, K. K. (1984) "Effect of preemphasis on vowel recognition performance," Speech Communication, vol. 3, pp. 101-106.

Paliwal, K. K., Ainsworth W. A. and Lindsay D. (1983) "A study of two-formant models for vowel identification," Speech Communication, vol. 2, pp. 295-303.

Pan, K. C., Soong, F. K. and Rabiner, L. R. (1985) "A vectorquantization based preprocessor for speaker-independent isolated word recognition," IEEE Trans. on ASSP, vol. 33, pp. 546-560.

Parthasarathy, S. and Tufts, D. W. (1987) "Excitation-synchronous modeling of voiced speech," IEEE Trans. on ASSP, vol. 35, no. 9, pp. 1241-1249.

Paul, D. B. and Martin, E. A. (1988) "Speaker stress-resistant continuous speech recognition," ICASSP-88, N.Y., S7.5, pp. 283-286.

Peterson G. E. and Barney, H. L.(1952) "Control methods used in a study of the vowels," JASA, vol. 24, pp. 175-184.

Pinson, E. N. (1963) "Pitch-Synchronous Time-Domain Estimation of Formant Frequencies and Bandwidths," JASA, vol. 35, pp. 1264-1273.

Pinto, N. B., Childers, D. G. and Lalwani, A. (1989) "Formant speech synthesis: Improving production quality," IEEE Trans. on ASSP, vol. 37, no. 12, pp. 1870-1885.

Pisoni, D. B. (1985) "Speech perception: Some new directions in research and theory," JASA, vol. 78, no. 1, Part 2, pp. 381-388.

Pister-Bourjot, C. and Haton, J. P. (1987) "Automatic learning: an approach to the adaptation of a speech recognition system to one or several speakers," Speech Communication, vol. 6, pp. 43-54.

Port, R. F., Reilly, W. T. and Maki, D. P. (1988) "Use of syllablescale timing to discriminate words," JASA, vol. 83, no. 1, pp. 265-273.

Rabiner, L. R. (1982) "Note on some factors affecting performance of Dynamic Time Warping algorithm for isolated word recognition," Bell System Technical J., vol. 61, no. 3, pp. 363-373.

Rabiner, L. R. (1978) "On creating reference templates for speaker independent recognition of isolated words," IEEE Trans. on ASSP, vol. 26, no. 1, pp. 34-42.

Rabiner, L. R. and Juang, B. H. (1986),, "An introduction to Hidden Markov Models," IEEE ASSP Magazine, pp. 4-16.

Rabiner, L. R. and Levinson, S. E. (1981) "Isolated and connected word recognition - theory and selected applications," IEEE Trans. on Communication, vol. 29, no. 5, pp. 621-659.

Rabiner, L. R. and Levinson, S. E. (1985) "A speaker-independent, syntax-directed, connected word recognition based on Hidden Markov Models and Level Building," IEEE Trans. on ASSP, vol. 33, no. 3, pp. 561-573 .

Rabiner, L. R., Levinson, S. E., Rosenberg, A. E. and Wilpon, J. G. (1979) "Speaker-independent recognition of isolated words using clustering techniques," IEEE Trans. on ASSP, vol. 27, pp. 336-349.

Rabiner, L. R., Levinson, S. E. and Sondhi, M. M. (1983) "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," Bell Syst. Technical J., vol. 62, no. 4, pp. 1075-1105.

Rabiner, L. R., Pan, C. K. and Soong, F. K. (1984) "On the performance of isolated word speech recognizers using vector quantization and temporal energy contours," AT&T Bell Labs Technical J., vol. 63, no. 7, pp. 1245-1260.

Rabiner, L. R., Rosenberg, A. E. and Levinson, S. E. (1975) "Considerations in dynamic time warping algorithms for discrete word recognition," IEEE Trans. on ASSP, vol. 26, no. 6, pp. 575-582.

Rabiner, L. R. and Sambur, M. R. (1977) "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," Proc. IEEE ICASSP, Hartford, CT, pp. 323-326.

Rabiner, L. R. and Schafer, R. W. (1978), Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L. R., Schmidt, C. E. and Atal, B. S. (1977) "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech," Bell System Technical J., vol. 56, no. 3, pp. 455-482.

Rabiner, L. R., Sondhi, M. M. and Levinson, S. E. (1984) "A vector quantizer combining energy and LPC parameters and its application to isolated word recognition," Bell Syst. Technical J., vol. 63, no. 5, pp. 721-735.

Rabiner, L. R. and Soong, F. K. (1985) "Single-frame vowel recognition using vector quantization with several distance measure," AT&T Technical J., vol. 64, no. 10, pp. 2319-2330.

Rabiner, L. R. and Wilpon, J. G. (1979a) "Considerations in applying clustering techniques to speaker independent word recognition," JASA, vol. 66, no. 3, pp. 663-673.

Rabiner, L. R. and Wilpon, J. G. (1979b) "Speaker-independent isolated word recogniton for a moderate size (54 word) vocabulary," IEEE Trans. on ASSP, vol. 27, no. 6, pp. 583-587.

Rao, C. R. (1973) Linear Statistical Inference and its Applications, 2nd ed., John Wiley & Sons, New York.

Reddy, D. R. (1976) "Speech recognition by machine: A review," Proc. IEEE, vol. 64, pp. 501-531.

Repp, B. (1986) "Perception of the [m]-[n] distinction in CV syllables," JASA, vol. 79, pp. 1987-1999.

Repp, B. (1987) "On the possible role of auditory short-term adaptation in perception of the pervocalic [m]-[n] contrast," JASA, vol. 82, pp. 1525-1538.

Repp, B., and Svastikula, K. (1988) "Perception of the [m]-[n] distinction in VC syllables," JASA, vol. 83, pp. 237-247.

Rissanen, J. (1978) "Modeling by shortest data description," *Automatica,* vol. 11, pp. 165-167.

Rosenberg, A. E. (1971) "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," JASA. vol. 49, no. 2, pp. 583-590.

Rosenberg, A. E. (1976) "Automatic speaker verification: a review," Proc. IEEE, vol. 64, pp. 475-487.

Rothenberg, M. R. (1973) "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," JASA, vol. 53, pp. 1632-1645.

Rothenberg, M. (1981) "Acoustic interaction between the glottal source and the vocal tract," Vocal Fold Physiology (ed. by K. N. Stevens and M. Hirano), Univ. of Tokyo Press, Tokyo, pp. 305-323.

Rumelhart, D.E. and McClelland, J.L. (1986) Parallel Distributed Processing, vol. 1, MIT Press, Cambridge MA.

Sakoe, H. and Chiba, S. (1978) "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. on ASSP, vol. 26, no. 1, pp. 43-49.

Sambur, M. R., Rosenberg, A. E., Rabiner, L. R., and McGonegal, C. A. (1978) "On reducing the buzz in LPC synthesis," JASA. vol. 63, pp. 918-924.

SAS (1988) SAS/STAT User's Guide, R6.03 ed., SAS Institute, Cary, NC.

Schroeder, M. R., and Atal, B. S. (1985) "Code-excited linear prediction (CELP): high quality speech at very low bit rates," Proc. IEEE ICASSP, Tampa, Florida, pp. 937-940.

Schroeter, J., Larar, J. N. and Sondhi, M. M. (1988) "Multi-frame approach for parameter estimation of a physiological model of speech production," ICASSP-88, N.Y., S2.6, pp. 83-86.

Seitz, P. F., McCormick, M. M., Watson M. C., and Bladon R. A. (1990) "Relational spectral features for place of articulation in nasal consonants," JASA, vol. 87, no. 1, pp. 351-358.

Seneff, S. (1982) "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction," IEEE Trans. on ASSP, vol. 30, no. 4, pp. 566-578.

Shirai, K. and Kobayashi, T. (1986) "Estimating articulatory motion from speech wave," Speech Communication, vol. 5, pp. 159-170.

Shore, J. E. and Burton, D. K. (1983) "Discrete utterance speech recognition without time alignment," IEEE Trans. on Information Theory, vol. 29, no. 4, pp. 473-491.

Smith, M. E., Robinson, K. E. and Strong, W. J. (1981) "Intelligibility and quality of linear predictor and eigenparameter coded speech," IEEE Trans. on ASSP, vol. 29, no. 3, pp. 391-395 .

Siegel, L.J. (1979a) "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier," IEEE Trans. on ASSP, vol. 26, no. 1, pp. 83-89.

Siegel, L.J. (1979b) "Features for the identification of mixed excitation in speech analysis," Proc. IEEE ICASSP, Washington D.C., pp. 752-755.

Siegel, L.J. and Bessey, A.C.(1982) "Voiced/unvoiced/mixed excitation classification of speech," IEEE Trans. on ASSP, vol. 29, no. 3, pp. 451-460.

Soderstrom, T. and Stoica, P. (1989) System Identification, Prentice-Hall, Inc., Englewood Cliffs.

Sondhi, M. M. (1975) "Measurement of the glottal waveform," JASA, vol. 57, pp. 228-232.

Soong, F. K. and Rosenberg, A. E. (1988) "On the use of instantaneous and transitional spectral information in speaker recognition," IEEE Trans. on ASSP, vol. 36, no. 6, pp. 871-879 .

Stanton, B. J., Jamieson, L. H. and Allen, G. D. (1988) "Acoustic phonetic analysis of loud and Lombard speech in simulated cockpit conditions," ICASSP-88, S8.7, pp. 331-334.

Steiglitz, K. (1977) "On the simultaneous estimation of poles and zeros in speech analysis," IEEE Trans. on ASSP, vol. 25, pp. 194-202.

Stevens, J. P. (1990), Intermediate Statistics: A Modern Approach, Lawrence Erlbaum Associates Publishers, Hillsdale, NJ.

Stevens, K. N. (1972) "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds," Proc. Seventh Int. Cong. of Phonet. Sci. (ed. by A. Rigault and R. Charbonneau), Mouton, The Hague, pp. 206-232.

Stevens, K. N. (1980) "Acoustic correlates of some phonetic categories," JASA, vol. 68, no. 3, pp. 836-842.

Stevens, K. N. and Blumstein, S. E. (1978) "Invariant cues for place of articulation in stop consonants," JASA, vol. 65, no. 5, pp. 1358-1368.

Streeter, L. A., Macdonald, N. H., and Galotti, K. M., (1983) "Acoustic and perceptual indicators of emotional stress," JASA, vol. 73, no. 4, pp. 1354-1360.

Strube, H.W. (1974) "Determination of the instant of glottal closure from the speech wave," JASA, vol. 56, pp. 1625-1629.

Strube, H. W. (1982) "Time-Varying Wave Digital Filters and Vocal Tract Models," Proc. IEEE ICASSP, Paris, France, pp. 923-926.

Suomi, K. (1984) "On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem," Speech Communication, vol. 3, pp. 199-209.

Syrdal, A. K. (1985) "Aspects of a model of the auditory representation of American english vowels," Speech Communication, vol. 4, pp. 121-135.

Syrdal, A. K. and Gopal, H. S. (1986) "A perceptual model of vowel recognition based on the auditory representation of American english vowels," JASA, vol. 79, no. 4, pp. 1086-1100.

Tardelli, J. D., Walter, C.M. . Sims, J.T, LaFollette, P.A.and Gatewood, P.D. (1986) "Research and development for digital voice processing," Rome Air Devel. Center Tech. Rep., vol. RADC-TR-86-171.

Tierney, J. (1980) "A study of LPC analysis of speech in additive noise," IEEE Trans. on ASSP, vol. 28, no. 4, pp. 389-397.

Ting, Y. T. (1989) *Adaptive estimation of time-varying signal parameters with applications to speech*, Ph.D. dissertation, University of Florida.

Ting, Y. T., and Childers, D. G. (1990) "Speech analysis using the weighted recursive least squares algorithm with a variable forgetting factor," Proc. IEEE ICASSP, Albuquerque, New Mexico, vol. 1, pp. 389-392.

Ting, Y.T., Childers, D.G. and Principe, J.C. (1988) "Tracking spectral resonances," IEEE Fourth Annual Workshop on Spectrum Estimation and Modeling, Minneapolis, Minnesota, pp. 49-53.

Tohkura, Y. (1987) "A weighted cepstral distance measure for speech recognition," IEEE Trans. on ASSP, vol. 35, no. 10, pp. 1414-1422.

Tou, J.T. and R.C. Gonzalez, (1974) Pattern Recognition Principles, Addison-Wesley, Reading, MA.

Un, C.K. and Lee, H.H.(1980) "Voiced/unvoiced/silence discrimination of speech by delta modulation," IEEE Trans. on ASSP, vol. 27, no. 4, pp. 398-407.

Vaissiere, J. (1985) "Speech recognition: A tutorial," Computer Speech Processing, ed. by Frank Fallside and William A. Woods, Prentice-Hall, Englewood Cliffs, NJ, pp. 191-242.

van Rossum N.J.T.M and Rietveld, A.C.M (1984), "A perceptual evaluation of V/U detector," Speech Communication, vol. 3, pp. 151-156.

van Veen, T. M. and Houtgast, T. (1985) "Spectral sharpness and vowel dissimilarity," JASA, vol. 77, no. 2, pp. 628-634.

Veeneman, D. E. and BeMent, S. L. (1985) "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Trans. on ASSP, vol. 33, no. 4, pp. 369-377.

Vysotsky, G. J. (1984) "A speaker-independent discrete utterance recognition system, combining deterministic and probabilistic strategies," IEEE Trans. on ASSP, vol. 32, no. 3, pp. 489-498 .

Wakita, H. (1977) "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE Trans. on ASSP, vol. 25, no. 2, pp. 183-192.

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A. and Schwartz, D. M. (1978) "Correlates of psychological dimensions in talker similarity," J. of Speech and Hearing Resear., vol. 21, pp. 265-275 .

Walley, A. C. and Carrell, T. D. (1983) "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," JASA, vol. 73, no. 3, pp. 1011-1022.

Welch, P. D. (1967), "The use of Fast Fourier Transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Trans. Audio and Electroacoust., vol. AU-15, pp. 70-73.

Wells, B.B. (1985), "Voiced/unvoiced decision based on the bispectrum," Proc. IEEE ICASSP, Tampa, FL, pp. 1589-1592.

Wong, C. (1991), *The implementation of glottal source-vocal tract interaction effects to improve the naturalness of synthetic speech*, Ph.D. dissertation, University of Florida.

Wong, D. Y., Markel, J. D. and Gray, Jr. A. H. (1979) "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Trans. on ASSP, vol. 27, no. 4, pp. 350-355.

Wu K. and Childers, D. G. (1990) "Gender recognition from speech: Part I. Coarse analysis and Part II. Fine analysis," submitted for publication.

Yegnanarayana, B., Saika, D. K. and Krishnan, T. R. (1984) "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," IEEE Trans. on ASSP, vol. 32, no. 3, pp. 610-623.

Yea, J.J. and Childers, D.G. (1983) "Detecting speech nasalization by a spectral based algorithm," ASSP Spectrum Estimation Workshop II, Tampa, Florida, pp. 84–88.

You-Han, Pao. (1988) Adaptive Pattern Recognition and Neural Networks. Case western Reserve University, Addition-Wesley Publishing Co.

Yumoto, E., Gould, W. and Baer, T. (1982) "Harmonic-to-noise ratio as an index of the degree of hoarseness," JASA, vol. 71, pp. 1544-1550.

Zue, V. W. (1985) "The use of speech knowledge in automatic speech recognition," Proc. IEEE, vol. 73, no. 11, pp. 1602-1615.

Zue, V. W. and Schwartz, R. M. (1980) "Acoustic processing and phonetic analysis," Trends in Speech Recognition (ed. by W. Lea), Prentice Hall, Englewood Cliffs, NJ, pp. 101-124.

Zwicker, E. and Terhardt, E. (1980) "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," JASA, vol. 68 no. 5, pp. 1523-1525.

BIOGRAPHICAL SKETCH

Mr. Kyosik Lee was born in Korea on Oct. 2nd, 1959. He graduated from the Kyongpook National University, Taegu, Korea, in February, 1982, with a B.Sc. degree in electronics engineering. He also received his Master of Engineering degree from the Kyongpook National University, Taegu, Korea, in February, 1984. He joined the Agency for Defense Development (ADD) as a researcher to be exempted from military service in March, 1984. He had been involved in research and development of the underwater acoustic weapon systems.

Since January, 1990, he has been a Graduate Research Assistant in Dr. D. G. Childers' Mind–Machine Interaction Research Center in the Department of Electrical Engineering at the University of Florida, Gainesville, Florida, where his primary area of interest is digital signal processing with application to speech analysis and synthesis techniques by computer. After completing the requirements for the Ph.D. degree, he expects to be a research member in the Electronics and Telecommunications Research Institutes in Korea.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Donald G. Childers, Chairman
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Leon W. Couch, II
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Fred J. Taylor
Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jose C. Principe
Associate Professor of Electrical Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Howard B. Rothman
Professor of Communication Processes
and Disorders

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December, 1992

Winfred M. Phillips
Dean, College of Engineering

Madelyn M. Lockhart
Dean, Graduate School